DOCUMENT RESUME

ED 434 136 TM 030 098

AUTHOR Salvucci, Sameena; Wenck, Stephen; Tyson, James
TITLE Development of a Prototype System for Accessing Linked NCES Data. Working Paper Series.
INSTITUTION Synectics for Management Decisions, Inc., Arlington, VA.
SPONS AGENCY National Center for Education Statistics (ED), Washington, DC.
REPORT NO NCES-WP-98-15
PUB DATE 1998-10-00
NOTE 77p.
AVAILABLE FROM U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics, 555 New Jersey Avenue, N.W., Room 400, Washington, DC 20208-5652; Tel: 202-219-1831.
PUB TYPE Reports - Evaluative (142)
EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS *Access to Information; Data Analysis; Elementary Secondary Education; *Information Dissemination; Models; *National Surveys
IDENTIFIERS *Linkage; *National Center for Education Statistics

ABSTRACT

A project has been developed to advance the capabilities of the National Center for Education Statistics (NCES) to support the dissemination of linked data from multiple surveys, multiple components within a survey, and multiple time points. An essential element of this study is the development of a software prototype system to facilitate NCES data customers' access to linked historical data (i.e., a data warehouse). Following an introduction, section 2 summarizes research in data warehousing. Section 3 documents the range of potential linkages using historical NCES survey data, describes recent linking projects, and explains the selection of data for the warehousing project. Section 4 addresses the objective of the data warehouse, explains metadata, and provides a description of the data model developed to represent multiple years of the Common Core of Data district level data set. Section 5 describes the steps and tools used in implementation of the model into a prototype warehouse, and section 6 contains conclusions based on this study. Section 7 contains selected references. Three appendixes contain supplemental information about the interviews, data imputation, and sample screens from the software. (Contains 5 tables, 3 figures, and 35 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

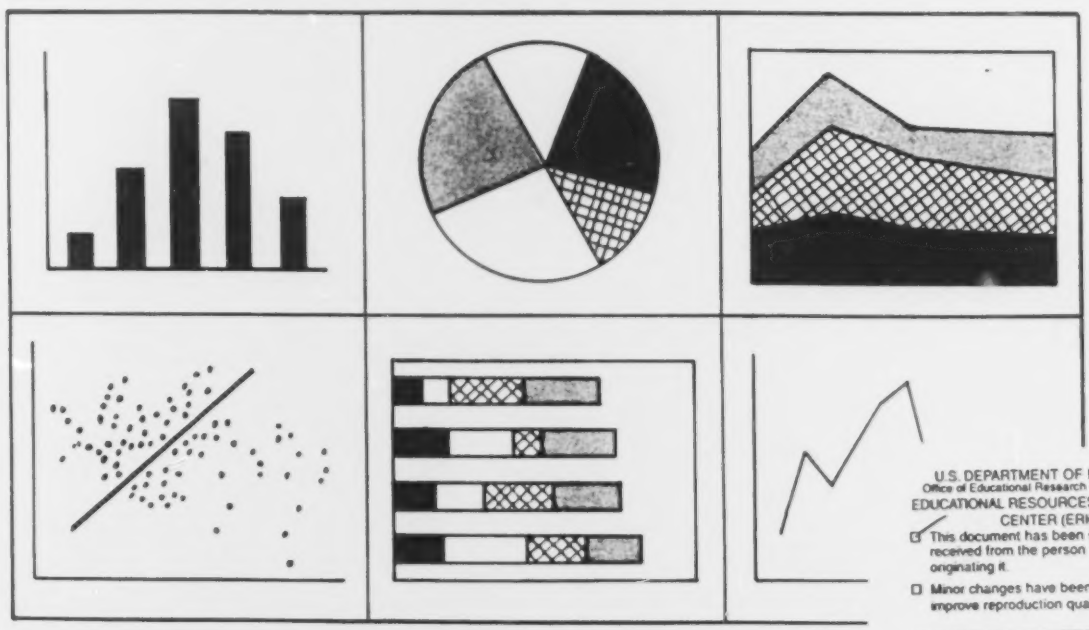
ED 434 136

NATIONAL CENTER FOR EDUCATION STATISTICS

Working Paper Series

Development of a
Prototype System for
Accessing Linked NCES Data

Working Paper No. 98-15 October 1998



U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM030098

U.S. Department of Education
Office of Educational Research and Improvement

BEST COPY AVAILABLE

Development of a Prototype System for Accessing Linked NCES Data

Working Paper No. 98-15 October 1998

Contact: Steven Kaufman
Surveys and Cooperative Systems Group
e-mail: steve_kaufman@ed.gov

U.S. Department of Education
Richard W. Riley
Secretary

Office of Educational Research and Improvement
C. Kent McGuire
Assistant Secretary

National Center for Education Statistics
Pascal D. Forgione, Jr.
Commissioner

The National Center for Education Statistics (NCES) is the primary federal entity for collecting, analyzing, and reporting data related to education in the United States and other nations. It fulfills a congressional mandate to collect, collate, analyze, and report full and complete statistics on the condition of education in the United States; conduct and publish reports and specialized analyses of the meaning and significance of such statistics; assist state and local education agencies in improving their statistical systems; and review and report on education activities in foreign countries.

NCES activities are designed to address high priority education data needs; provide consistent, reliable, complete, and accurate indicators of education status and trends; and report timely, useful, and high quality data to the U.S. Department of Education, the Congress, the states, other education policymakers, practitioners, data users, and the general public.

We strive to make our products available in a variety of formats and in language that is appropriate to a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other NCES product or report, we would like to hear from you. Please direct your comments to:

National Center for Education Statistics
Office of Educational Research and Improvement
U.S. Department of Education
555 New Jersey Avenue, NW
Washington, DC 20208

The NCES World Wide Web Home Page is
http://nces.ed.gov

Suggested Citation

U.S. Department of Education. National Center for Education Statistics. *Development of a Prototype System for Accessing Linked NCES Data*. Working Paper No. 98-15, by Sameena Salvucci, Stephen Wenck, and James Tyson. Project Officer, Steven Kaufman. Washington, D.C.: 1998.

October 1998

Foreword

Each year a large number of written documents are generated by NCES staff and individuals commissioned by NCES which provide preliminary analyses of survey results and address technical, methodological, and evaluation issues. Even though they are not formally published, these documents reflect a tremendous amount of unique expertise, knowledge, and experience.

The ***Working Paper Series*** was created in order to preserve the information contained in these documents and to promote the sharing of valuable work experience and knowledge. However, these documents were prepared under different formats and did not undergo vigorous NCES publication review and editing prior to their inclusion in the series. Consequently, we encourage users of the series to consult the individual authors for citations.

To receive information about submitting manuscripts or obtaining copies of the series, please contact Ruth R. Harris at (202) 219-1831 (e-mail: ruth_harris@ed.gov) or U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics, 555 New Jersey Ave., N.W., Room 400, Washington, D.C. 20208-5654.

Marilyn McMillen
Chief Statistician
Statistical Standards and Services Group

Samuel S. Peng
Director
Methodology, Training, and Customer Service Program

Development of a Prototype System for Accessing Linked NCES Data

Prepared by:

Sameena Salvucci
Stephen Wenck
James Tyson

Synectics for Management Decisions, Inc.

Prepared for:

U.S. Department of Education
Office of Educational Research and Development
National Center for Education Statistics

October 1998

Table of Contents

1. INTRODUCTION 1

2. RESEARCH TOWARD DEVELOPING A PROTOTYPE 3

2.1. COMMERCIAL-OFF-THE-SHELF (COTS) PRODUCTS AND TECHNOLOGIES 3
2.2. SIMILAR ACTIVITIES BY OTHER GOVERNMENT ORGANIZATIONS 4
2.2.1. The National Agricultural Statistics Service (NASS) 5
2.2.2. The Environmental Protection Agency (EPA) 5
2.2.3. The Bureau of Labor Statistics (BLS) 6
2.2.4. The Census Bureau 6
2.2.5. The National Science Foundation (NSF) 7
2.2.6. Statistics Canada 7
2.3. REVIEW OF INDUSTRY LITERATURE 8
2.4. CONCLUSIONS 8

3. NCES DATA LINKAGES 9

3.1. POTENTIAL LINKAGES 9
3.2. RECENT SURVEY LINKING PROJECTS 12

4. PROTOTYPE WAREHOUSE: METADATA AND THE DATA MODEL 15

4.1. METADATA 15
4.2. DATA MODEL 16

5. PROTOTYPE WAREHOUSE: IMPLEMENTATION 19

5.1. ORACLE WAREHOUSE 21
5.2. COGNOS TOOLS 22

6. LESSONS LEARNED AND POSSIBLE NEXT STEPS 24

7. REFERENCES 25

APPENDIX A: SUMMARY OF INTERVIEWS 29
APPENDIX B: LONGITUDINAL EDITING AND IMPUTATION OF CCD DATA 45
APPENDIX C: SAMPLE COGNOS SCREEN SHOTS 51

1. Introduction

In the past 20 years, there has been a significant change in the way data are accessed and used. The procedural tools of the seventies have given way to graphical user interfaces (GUIs) which allow direct manipulation of data. This evolution has created an opportunity to place powerful tools for information retrieval and manipulation in the hands of users.

NCES has been a catalyst in the development and innovative use of technology, including the addition of user tools with its survey data. It fueled advances such as the electronic code book (ECB), the data analysis system (DAS), and other data products including the Common Core of Data (CCD) and the Integrated Postsecondary Education System (IPEDS) CD-ROMs with graphical user interfaces. These were not just technical advances; they provided a paradigm shift from merely providing information to placing user-friendly tools along with the data directly in the hands of users in order to improve the accessibility and usefulness of the information.

Recently, NCES began providing its survey data and tools through the Internet, a step which has greatly expanded accessibility to these data. However, both NCES internal staff and its customers have a further need for the ability to directly access "linked" data from multiple surveys, multiple components within a survey, and multiple time points, for methodological and analytic purposes.

The purpose of this project is to advance NCES' capabilities to support the dissemination of linked data[1]. This enhanced capability acknowledges the critical role of providing access to linked historical data in improving analytical capabilities, improving sampling and estimation techniques, ensuring data quality, and improving customer service. An essential element of this study is the development of a software prototype system to facilitate NCES data customers' access to linked historical data (i.e., a data warehouse).

The basic concept of an NCES data warehouse is to facilitate data management such that specific information is easily accessible to all users of NCES data.

The major expected benefits to NCES of such a data warehouse system are:

- Elimination of the need to re-create links every time a research purpose requires it, thereby substantially reducing the effort involved in ad-hoc survey linkages;

- Wider dissemination and use of its survey data through more user friendly access to multiple-linked data products;

[1] "Data linking" and "linked data" refer to the concept of associating data from *across* various survey data sets through common or related elements. This may include linking heterogeneous data sets (such as via common data elements), as well as linking homogeneous data sets (such as to aggregate results across different survey years), or both.

- Establishment of "standards" for documentation of data and metadata related to future releases of NCES data.

The initial steps in the development of the prototype include research in the capabilities of available tools and techniques for building data warehouses, a review of prominent examples of warehouse implementations in the federal agencies, and a review of types of historic and ongoing NCES data linking activities.

This report has six main sections. Section 2 summarizes the findings of our research in the area of data warehousing. Section 3 documents the range of potential linkages using historical NCES survey data, briefly describes recently completed survey linking projects, and explains the selection of survey data to be used in the data warehouse prototype. Section 4 addresses the objective of the prototype warehouse, explains metadata, and provides a description of the data model developed to represent multiple years of the Common Core of Data (CCD) district level data set. Section 5 describes the steps and tools used in the implementation of the data model into a prototype warehouse. Section 6 provides a set of conclusions based on this study. Section 7 includes a selected set of references.

2. Research Toward Developing a Prototype

This section summarizes the research findings for this project. The research spanned three basic areas that relate to NCES' interest in data warehouses:

- Identifying and comparing commercial off-the-shelf products and technologies;
- Reviewing activities by other organizations with similar mission and scope; and
- Reviewing white papers, articles and textbooks.

The activities and findings in each of the areas of primary research undertaken are summarized below.

2.1. Commercial-off-the-shelf (COTS) products and technologies

Investigation of COTS products focused on 1) characterizing the broad range of vendors and products that are positioned to support the general goals of this project, and 2) evaluating selected products in somewhat closer detail towards the development of a data linking prototype.

The current market for COTS software products that support data warehousing is at a stage of rapid growth. Many competing products from numerous vendors characterize the current market. One Internet-based list of data warehousing products (maintained by Larry Greenfield, LGI Systems at http://pwp.starnetinc.com/larryg/index.html) recently identified 822 products in 17 categories:

Product category	No. of products
Report and Query	118
OLAP / Multidimensional Databases	60
Executive Information Systems	53
Data Mining	100
Document Retrieval	66
Geographic Information Systems	29
Decision Analysis	29
Process Modeling	24
Statistics	39
Information Filtering	26
Other End User Decision Support Tools	20
Data Extraction, Cleaning, Loading	139
Information Catalogs	14
Databases for Data Warehousing	30
Query and Load Accelerators	23
Middleware	33
Other Database Tools	19

The first two categories of products—that is, Report and Query tools, and OLAP tools—were determined to be the most appropriate types of tools to use to demonstrate the capabilities of linked survey data.

Report and query tools produce tabular reports with simple summations and aggregations, typically (but not necessarily) based on the contents of a relational database. Examples of popular products with this capability include Cognos Impromptu, Microsoft Access, Brio BrioQuery, Seagate Software Crystal Reports, and the SAS System.

On line analytical processing (OLAP) is a popularly used to describe an interactive approach to decision support[2]. OLAP query tools produce reports with more complex processing requirements, and typically work against (star-schema[3]) relational databases and/or multidimensional databases (such as Red Brick or Essbase). Examples of popular products with this capability include Cognos PowerPlay, Brio BrioQuery, IQ Software IQ/Vision, Seagate Software Crystal Info., and the SAS System.

Note that there is not a clear line between the two categories of products (e.g., Cognos, Brio and SAS products fit in both). Generally, OLAP means more complex processing than Report and Query. Just about EVERY product in either category runs under Windows and can connect to a relational database management system (RDBMS) such as Oracle or SQL Server.

Selected vendors (Brio Technology, SAS Institute, Information Builders, Inc. and Cognos) were chosen for on-site review of their product offerings. Vendors of software specifically purported to support linking statistical data sets across time were also contacted. For prototype purposes, Report and Query and On-Line Analytic Processing (OLAP) software packages, Cognos's Impromptu and PowerPlay, were chosen.

2.2. Similar activities by other government organizations

The activities of several Federal government agencies with missions similar to NCES were reviewed, including the National Agriculture Statistics Service, the Environmental Protection Agency, the Bureau of Labor Statistics, the Census Bureau, and the National Science Foundation.

General findings show that these agencies are taking or have taken an approach to linking statistical data similar to NCES' current approach, that is to plan and implement various data warehousing technologies to support the efficient organization and dissemination of data and metadata.

---

[2] Decision support activities involve producing reports and views of aggregated data, such as cross tabulations and various statistical measures to support inferential decision making.

[3] Star Schema is a name that database designers have used to describe dimensional models because the diagram of this type of model looks like a star, with one large central table and a set of smaller attendant tables displayed in a radial pattern around the central table.

2.2.1. The National Agricultural Statistics Service (NASS)

The National Agricultural Statistics Service (NASS) of the Department of Agriculture has developed a relational database system under the direction of Mickey Yost. This relational database employs the star schema model. NASS has chosen to employ the Red Brick Warehouse management system. NASS has selected Brio as their user interface software package. They are also looking into developing a SAS CONNECT user interface.

In particular, the Department of Agriculture has successfully developed a system that links various survey data sets (including time series data) using a multidimensional or star-schema model. The Department of Agriculture staff who developed this system emphasized the practical value of the star-schema data model as their starting point, particularly as the use of a common model allows them flexibility in the choice of COTS products that can support or work with their data.

By design, the NASS warehouse is not available on the web. Their warehouse contains restricted data and is only available on the NASS LAN. The warehouse currently only supports the Brio OLAP tool. By design, users can print Brio crosstabs in a number of formats (ASCII, Excel, etc.), but cannot export raw data.

2.2.2. The Environmental Protection Agency (EPA)

EPA has developed an on-line (web based) relational database called Envirofacts. This database integrates data extracted monthly from five facility (site) based EPA program systems. Those program systems are: Superfund Data, Hazardous Waste Data, Toxic Release Inventory, Water Discharge Permits, and an Aerometric Information Subsystem (AIRS). Envirofacts also contains a grant information database, three integrating databases, and mapping applications. The Envirofacts database contains only data available under the Freedom of Information Act and therefore full access is granted to all users.

Envirofacts has been created by extracting data from the mainframe computer versions of these data sources and placing it in an ORACLE Relational Database Management System. It is updated monthly. Envirofacts can be accessed by any software package that can connect to an ORACLE database. Information in the Envirofacts database can be freely accessed through the use of predetermined or user-developed queries. While the complexity of user-developed queries is unlimited, queries that return a large volume of data may terminate prematurely due to system limitations. The constraints established for the enviro user account are:

- Each query is limited to 2.5 minutes of CPU time.
- A single session can be 15 minutes in duration.

- A session may stay idle for 7.5 minutes, at which time it will be terminated.

Users are urged to develop queries that return small batches of data and terminate their session as soon as their queries have executed, so that others may be able to access the database.

The Envirofacts Query allows users to retrieve the environmental profile of facilities that match the query specifications. Queries are allowed on the following criteria: facility name, geography, or Standard Industrial Classification (SIC) Code. The profile of facilities includes information regarding toxic chemical releases, chemical permit compliance, hazardous waste handling processes, Superfund status, and air emission estimates for pollutants regulated under the Clean Air Act. Query Mapper has been developed to map the results of Envirofacts queries.

2.2.3. The Bureau of Labor Statistics (BLS)

The Bureau of Labor Statistics (BLS), Department of Labor has developed a web-based data warehouse. The data are not linked across different surveys. The system allows extraction of timeseries data for a specific database. The only output selection is raw ASCII data. Within each database one can select any one or all of the timeseries shown and choose desired date ranges and output options. There are also a limited number of time series data for which BLS allows the requester to select a subset of the available variables to be output as ASCII data. The users would then transfer these ASCII data files into their choice of statistical packages (e.g., SAS, SPSS) for analysis.

2.2.4. The Census Bureau

The Census Bureau has developed the "Data Extraction System" (DES) as a web-based data warehouse. The Data Extraction System is a tool for extracting records (rows) and fields (columns) from very large, public-information, data files (for example: survey and census records.) The system produces custom extracts in selectable data file formats which, when processed, can then be picked up at the Census Bureau's FTP site. This system does not produce tabulations, it produces only raw data which must then be processed by programming or statistical analysis software elsewhere. No confidential data are available via this service. As with the BLS system, data are not linked across databases, but rather allow for extraction of data from a particular data source. The DES requires that the user specify the survey, the number of records, and the survey items and returns an e-mail to the user when the requested ASCII data are available for downloading. The users would then need to transfer the ASCII data file into their choice of statistical packages (e.g., SAS, SPSS) for analysis.

2.2.5. The National Science Foundation (NSF)

The National Science Foundation has designed WebCASPAR, a database system to provide quick and convenient access to a wide range of statistical data focusing on U.S. universities and colleges and their science and engineering resources. WebCASPAR users can specify the statistical data of interest and either view the data through their web browsers or transfer the data to their own computers as spreadsheets or other data files.

2.2.6. Statistics Canada

For years, Statistics Canada has created the E-STAT CD-ROM to disseminate their socio-economic databases to the national school community. This offering is now available to that community through SchoolNet on the World Wide Web.

The data bases contained in E-STAT are:

- Census Profiles showing various characteristics across detailed geographic areas; and
- the CANadian Socio-economic Information Management (CANSIM) data base containing over 200,000 time series of socio-economic data detailing numerous subject matter areas at various levels of geography.

The databases are stored in a proprietary data base format. Users can get the data they want in two primary ways: hypertext (HTML) navigation through subject matter areas, and by keyword searching.

Once the desired data has been specified, a request is uploaded to the WebServer and the data retrieval is executed. The user is then asked which of several formats is desired (e.g. HTML, Comma-Separated Value, Ivation's Beyond 20/20, etc.). On selection of the desired format, the server will then download the selected data in the selected format to the user's workstation.

Users who want to view and manipulate tables in Beyond 20/20 format need to have the Beyond 20/20 Data Browser. The Data Browser is available on the E-STAT WebServer in a file, which is ready for installation once downloaded. Clients can download the Browser once and subsequently use the Browser to "slice and dice", make selections, create charts and maps based on any table that is downloaded from the server thereafter.

2.3. Review of industry literature

More than 30 recent industry articles, white papers and texts on the subjects of linking data and data warehousing were identified and reviewed. Section 7 contains a partial bibliography.

A common theme of the literature was the importance and value of developing appropriate data models to represent the information that is to be linked or warehoused. While there was some debate over the relative benefits of alternative technologies for implementing these models, there was broad and strong consensus that the so-called dimensional or star-schema data model best represents data and metadata[4] in forms useful to analysts.

It should be noted that the multidimensional or star-schema model is a *conceptual* model, and theoretically can be implemented through any of a variety of *physical* designs. For example, there is an ongoing debate over the relative merits of general-purpose relational database management systems (RDBMS) versus special-purpose "multidimensional database systems" (MDDSs) towards supporting star-schema warehouses. Such discussions are important for developing an architecture to support production operations but, for purposes of this prototype, it was determined that a standard RDBMS already supported at NCES would be adequate to demonstrate the feasibility and utility of linking NCES data sets through a star-schema model.

2.4. Conclusions

In the course of our research, we refined our understanding of the critical elements of the prototype, i.e., to identify an approach to linking data sets that is flexible and scaleable; that has the potential for integration with WWW-based services; and that is appropriate to NCES' operating environment and goals. The research also indicated that a primary step in creating the prototype should be the development of an appropriate "dimensional" data model.

General reasons for this conclusion include:

- The literature reviewed almost universally recommended the development of an appropriate data model as a starting point.
- Data warehousing and, specifically, data linking activities by other organizations are generally organized around a specific data model.
- The commercial products evaluated all support a dimensional data model.

[4] Metadata is a general term for data about data.

3. NCES Data Linkages

This section documents a range of potential linkages using historical NCES survey data, briefly describes recently completed survey linking projects, and explains the selection of survey data to be used in the data warehouse prototype.

3.1. Potential Linkages

As a first step toward selecting the subset of NCES survey data to use in the development of the prototype data linking system, a review of NCES data sets was conducted to determine the range of potential linkages. For purposes of this report, linkage will refer to three methods of combining statistical information, including the ability to:

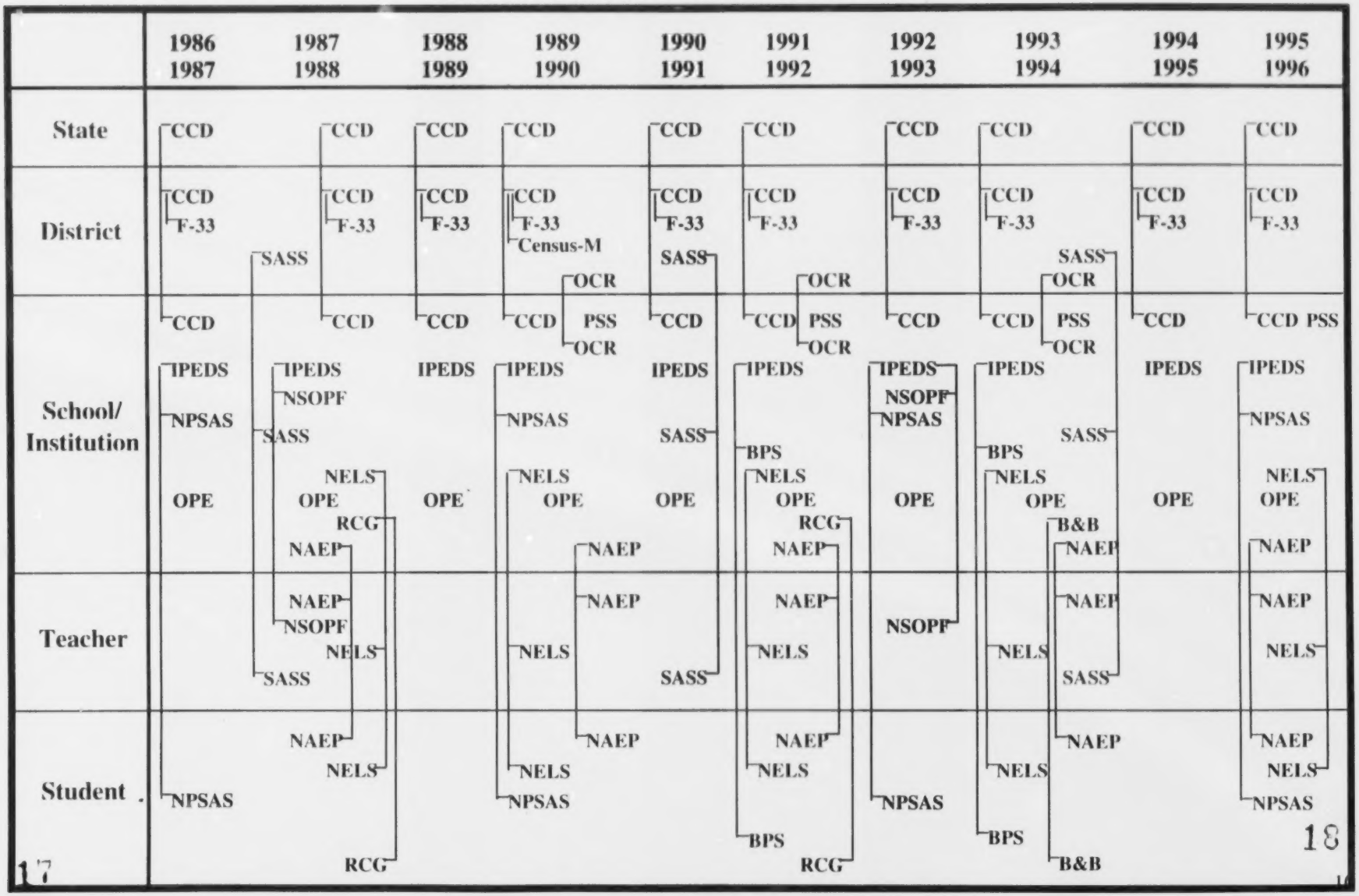
- Match identical entities over time within the same survey.
- Match identical entities across different surveys or survey components for a given point in time.
- Combine or compare statistics collected from different surveys, at some level of aggregation greater than an individual respondent unit.

With the assistance of NCES, all such potential linkages between the major NCES universe and sample survey data files were identified. Figure 3.1 illustrates the extent of potential linkages between survey files across time and across each of five levels of data aggregation (state, district, school/institution, teacher, and student) in a matrix form.

Each cell in the matrix contains the names of the set of surveys that were conducted at a particular level of aggregation in a particular year. For example, the information in the cell in the first row and first column indicates that the Common Core of Data (CCD) was collected at the state level in 1986-87. The lines connecting two surveys within a cell in the matrix indicate the ability to match identical entities across different surveys within a particular year and level of aggregation. For example, the line connecting CCD and F-33 in the second row and first column in the matrix indicates that the 1986-87 CCD and the 1986-87 F-33 (School District Finance Survey) are linkable at the district level.

The lines connecting the survey across cells in the matrix indicate the ability to match identical entities across different surveys/survey components within a particular year. For example, the line connecting CCD in the first row and first column to the CCD in the second row and first column indicates the ability to match the school districts in the 1986-87 CCD with the corresponding state in the 1986-87 CCD. This type of linkage provides the ability to compare statistics at the state level from the CCD school district level file (aggregated to the state level) to the CCD state level statistics. However, note that for sample survey data, even if the data were collected at a particular level, the estimated statistics may not be valid at that level of aggregation. For example, even though the 1987-88 Schools and Staffing Survey (SASS) has a component that collects data at the district level, the statistics are only valid when estimated at the national level.

Figure 3.1--Potential NCES Survey Linkages



BEST COPY AVAILABLE

Although the matrix does not include lines connecting surveys across years, this is another type of potential linkage. For example, the fact that CCD appears in all the cells in the first row of the matrix indicates that there is the ability to match CCD files across all those years at the state level.

Surveys that are not included in this matrix include surveys such as the Fast Response Survey System (FRSS) and the Postsecondary Education Quick Information System (PEQIS). These surveys have changing scope over time and therefore do not lend themselves to this type of analysis.

A list of the full names of each of the survey abbreviations indicated in figure 3.1 are provided below:

Abbreviation	**Survey Name**
Universe Surveys:	
CCD	Common Core of Data
IPEDS	Integrated Postsecondary Education Data System
PSS	Private School Survey
F-33	School District Finance Survey[5]
Sample Surveys:	
B&B	Baccalaureate and Beyond
BPS	Beginning Postsecondary Survey
Census-M	School District Data Book (Census Mapping Project)
OCR	Office of Civil Rights Survey
OPE	Office of Postsecondary Education Survey
NELS	National Educational Longitudinal Survey
NAEP	National Assessment of Educational Progress
NSOPF	National Study of Postsecondary Faculty
RCG	Recent College Graduate Survey
SASS	Schools and Staffing Survey

[5] These data are collected annually through the Bureau of Census' F-33, Survey of Local Governments. Data are collected from all districts in the census year (e.g., 1990) and in years ending in 2 and 7, and from a large sample in remaining years.

3.2. Recent survey linking projects

Based on the documentation of potential NCES data linkages, and the review of a range of projects that have required linking NCES data, it was decided that the prototype development process should take advantage of the results of previous linking efforts. In other words, the data sets and technical documentation produced in one of these linking efforts would be used as the data/metadata for the prototype. The decision on which linking activity to use was based on information collected through interviews conducted with both NCES staff and its contractor staff. The interviews included questions about the type of data sets that had been linked, the software that was used to perform the linkage, and the current availability of both the linked data and the corresponding documentation (including program code) on the linkage. Summaries of the range of data linking activities uncovered are provided in tables 3.1, 3.2 and 3.3. Table 3.1 includes information on the linkages involving CCD data and table 3.2 includes information on the linkages involving IPEDS data. Other linking projects are summarized in table 3.3. More detailed descriptions of the interview findings are provided in Appendix A.

The criteria used for reviewing these linking activities included the recency of the linking activity, the quality of the analysis performed (e.g., the results were adjudicated), and the existence of both the linked data set and corresponding documentation. Based on these criteria, a decision was made to use a recently completed linking activity carried out by Don McLaughlin at the American Institutes for Research (AIR) who used multiple years of district level CCD data. His analysis was done across regular school districts between the years 1986-87 and 1991-92. The linking activity included performing longitudinal editing and imputation of the data related to regular school districts in the CCD Public Education Agency Universe files for these years. See Appendix B for details of the steps taken to edit and impute the CCD data.

Table 2.1 -- Linking projects involving CCD

Surveys Linked	Name	Organization	LinkedOn	Variables Used	Documentation	Data Available?
68-92 CCD & 68-84 OCR	Mike Ross	NCES	District	Civil Rights Indicators	Not available	No
86-87 CCD & 93-94 NAEP & SASS State Assessment	Don Mclaughlin	AIR	District & State	Enrollment & Special Ed	Report as Appendix (Lee Hoffman, NCES) Programs	No
CCD & SASS (3 Waves) & FBI UCR & Climatic File & College Quality & City, County Data Book & Census	Jay Chambers	AIR	District & School	Wages & FBI Crime by City and Climate	Technical Report "Patterns of Teacher Compensation"	No
CCD & SASS	David Figlio	Univ. of Oregon	District	Cost	Not available	No
CCD & NELS	Bill Fowler	NCES	School	Achievement & Expenses & SES (Control)	A Report & SAS programs	No
CCD & SASS / NSOPF & IPEDS	Valerie Conley	Synectics	Faculty & Institution & Schools & Districts & States	Not available	CCD Net Users Manual	No
CCD & OCR & Census	Bill Sonnenberg	NCES	District	Not available	Not available	No
CCD Public School Universe & CCD Public Education Agency Universe	John Sietsema	NCES	District & Schools	Not available	Trends in School Demographics	No
CCD & SASS & Census & F33	Frank Johnson	NCES	District & School	Finance	Documents in each program	No
CCD& SASS & QED & IPEDS	Steve Owens	Census	District & School	Integrated Sampling Project	Draft Report available from Jonaki Bose (NCES)	No
CCD& SASS & Census & USDA Beal Code	Mike Podgursky	Univ. of Missouri	District & State	Graduation & Cost of Living & Salaries & Median Home Prices	Not available	No
CCD& SASS & IPEDS	Bob McArthur	Census	State	Finance	Not available	No
CCD & SASS	Arnold Reznek	Census	Schools & Districts	Not available	Programs	No
CCD & SASS	Sameena Salvucci	Synectics	Schools, Districts, & States	Enrollment, # of teachers, # of students,# of schools, # of LEAs	Administrative Records and Sample Survey Comparisons	Yes
CCD & SASS	Fan Zhang	Synectics	Schools	Enrollment, school level, full-time teachers, ethnic origin, etc.	CCD Adjustment to the SASS: A Comparison of Estimates	Yes

BEST COPY AVAILABLE

Table 2.2 -- Linking projects involving IPEDS

Surveys Linked	Name	Organization	LinkedOn	Variables Used	Documentation	Data Available?
IPEDS & NSOPF	Linda Zimbler	NCES			Representative of sample	Yes
IPEDS & NPSAS & B and B	Paula Knepper	NCES	Institution	Enrollment Degrees	Representative of sample	Yes
IPEDS & NHES	Kathryn Chandler	NCES	Adult Ed. District Early Childhood	Discipline household	Manual codebooks	No
IPEDS & NPSAS	Drew Malizio	NCES	District	Attendance Costs	No formal docs.	No
IPEDS & NELS	Sam Barbett	NCES	District	Enrollment Finance	Ad hoc programs	No
IPEDS & RCG	Peter Stowe	NCES	Unit	Employment	Ad hoc"not high quality"	No

Table 2.3 -- Other linking projects

Surveys Linked	Name	Organization	LinkedOn	Variables Used	Documentation	Data Available?
SDDB & SDAB	Larry McDonald	Census	District	Finance	Not available	No
NELS	Dale Ballu	University of Massachusetts	District	Private Schools, County ID, Zipcode	Not available	No
SASS & QED	Sameena Salvucci	Synectics	School, District	# of students, # of teachers	Report	Yes

BEST COPY AVAILABLE

4. Prototype Warehouse: Metadata and the Data Model

This section introduces the specific objective of the prototype warehouse, explains metadata as it pertains to the prototype, and provides a description of the dimensional model developed to represent multiple years of the Common Core of Data (CCD) district level data set.

The objective of this prototype warehouse is to provide a highly interactive *ad hoc* analysis system with the ability to access data from multiple NCES surveys simultaneously. Users expect to view this data from different perspectives – enrollment by race/ethnicity by year, number of full-time equivalent teachers by urbanicity of school by region, etc. – and expect to switch interactively among these perspectives. Users need to see information at different levels of detail, looking for insights with summary data (e.g. enrollment by race/ethnicity by year), then "drilling down" to increasing levels of detail (e.g. enrollment by race/ethnicity by school district by year), in order to understand root causes and anomalies.

Since the prototype will only include universe data (i.e., CCD), as described earlier, the ability to provide weighted data and correct standard errors was not an objective of this prototype warehouse system.

4.1. Metadata

Metadata is data about data. There are two categories of metadata: technical and warehouse-dependent. Technical metadata is the description of the data needed by various tools to store, manipulate, or move data. These tools include relational databases, application development tools, and OLAP tools just to name a few. Warehouse-dependent metadata, on the other hand, is the description of the data needed by users to understand the context and meaning of the data. It is any information that makes the data more useable by the researchers.

Building data warehouses is resource-intensive, so it is important to make the results of these projects usable by researchers. Implementing full warehouse-dependent metadata produces the significant benefit of making the information visible, understandable, and available. In short, it can be the difference between success and failure of a warehouse effort.

The warehouse-dependent metadata in the NCES prototype system is described in the next section. Essentially all categorical variables were put into the dimension tables. The only categorical variables not put into dimension tables were the imputation flags and the CCD question items. The imputation flag information was added in later version of the prototype. However, the CCD question items were not added to the prototype due to a limitation of the tool we chose. Cognos can give a general explanation of a variable for the whole warehouse, but if that information changes from year to year or from survey to survey, then Cognos currently can not handle that information.

4.2. Data Model

A data model provides an *abstract* view of the data, including a description of *what* data are actually to be stored in a database, and the relationships that exist among the data. In particular, the dimensional or star-schema data model is a logical design technique that seeks to present the data in a standard, intuitive framework that allows for high-performance access in a data warehouse.

A star-schema model was developed for the selected longitudinal CCD LEA data. See figure 4.1. The star schema represents to the end user a simple and query-centric view of the data by partitioning the data and the warehouse-dependent metadata into two types of tables: fact tables and dimension tables.

In the NCES prototype, fact tables represent the actual data from the multiple CCD data sets. The most useful facts in the fact tables are numeric and additive. Additivity is crucial because data warehouse application users almost never retrieve a single fact table record; rather, they fetch back hundreds and thousands of these records at a time, and the only useful thing to do with so many records is to add them up and present the information in an aggregate form. Information in the dimensions is what describes the facts.

Dimension tables, by contrast, most often contain descriptive textual information including the warehouse-dependent metadata. Dimension attributes are used as the source of most of the interesting constraints in data warehouse queries, and they are virtually always the source of the row and column headers in the users answer set. The power of the warehouse database is proportional to the quality and depth of the dimension tables.

A fact table has columns of anonymous index keys, one for each dimension, that relate back to exactly one row in each of the dimension tables. The variables listed in the fact table in figure 4.1 labeled FK (for foreign key) are these keys. The unique combination of keys in the fact table acts as a cross-reference specifying the intersection of each dimension at that data point. Examples of the types of facts pulled from the CCD datasets are also listed in this table.

The creation of unique dimension table keys is very important. These keys have to be unique to identify each record in the dimension table, but they also have to be generic enough so that as other surveys are added, a completely new key does not have to be added.

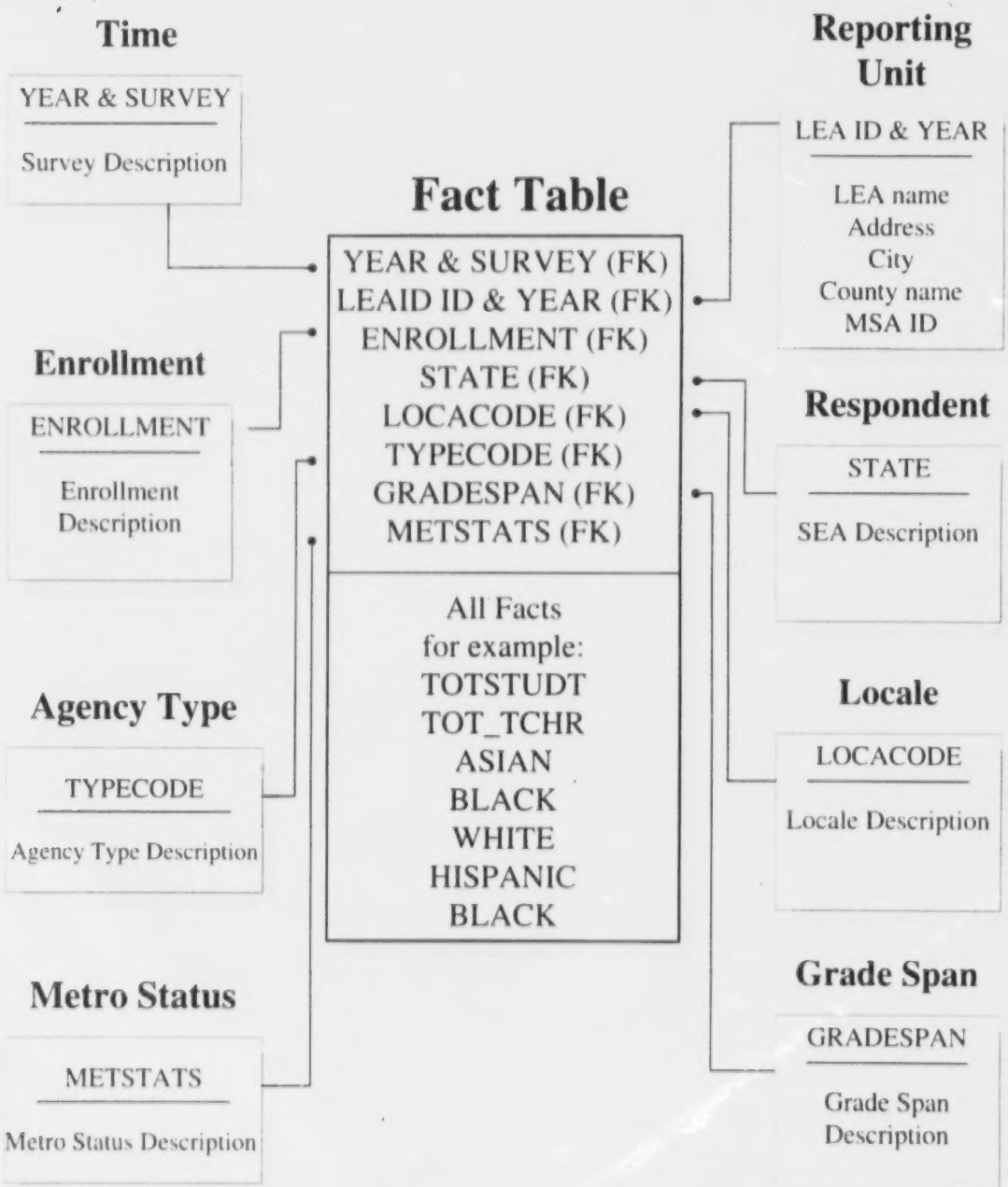
Because the single key in a dimension table is a row, metadata about that key can exist on that row and be available for querying. In fact, the dimensional intersection specified by the unique combination of foreign keys in the fact table defines the data point by presenting the metadata and the data item together.

Essentially, all categorical variables were put into the dimension tables. The only categorical variable not included as a dimension was the information on the imputation flags because of difficulty with the interpretation of the codes. The NCES dimension tables (figure 4.1) include Time, Enrollment, Agency Type, Reporting Unit, Respondent, Locale, and Grade

Span. The Time dimension includes the description of the survey year and survey name (in the case of the prototype this is just CCD LEA Universe but could be expanded to include other components of CCD or other sample or universe surveys at NCES) and a longer survey description. The Enrollment dimension includes the descriptions corresponding to each enrollment level code of the district. The Agency Type dimension includes the description corresponding to each agency type code in the CCD. The Metro Status dimension includes the description corresponding to each metro status code of the district. The Reporting Unit dimension provides each of the LEA names and corresponding address, city, county name, and MSA ID while the Respondent dimension includes each of the state education agency descriptions. Finally, the Grade Span dimension includes the description corresponding to each of the grade span codes for the district.

Figure 4.1

Dimensional Model for the NCES Warehouse Prototype Using 1986-1993 Longitudinal CCD LEA Data



5. Prototype Warehouse: Implementation

This section describes the steps and tools used in the implementation of the data model into a prototype warehouse.

The approach on the implementation of the data-linking prototype was to use standard methods and tools that facilitate end user access to linked or linkable NCES data. Concentrating on standard methods and tools ensures that the successful elements of the data-linking prototype can readily be integrated or incorporated within broader NCES or Department of Education plans, standards and technologies.

The prototype development process included the following steps:

1. Development of a flexible data model for the target data using an industry standard approach (i.e., the "dimensional" or "star schema" model). Dimensional modeling is a technique for visualizing the data as a "cube" of three, four, or even five or more dimensions. This can be illustrated with a simple example using the CCD:

   "States report data every year on district enrollment and locale type"

For the data warehouse design, special emphasis is added as follows:

   "*States* report data *every year* on enrollment and *locale type*"

   Most people find it easy to think of this as a **cube** of data, with labels on each of the edges of the cube. For the description above, the edges of the cube can be labeled as *State, Time, and Locale Type*. The points inside the cube are where the enrollment measurements for that combination of State, Time, and Locale Type are stored. This is the dimensional model. Star Schema is a name that database designers have used to describe dimensional models because the diagram of this type of model looks like a star, with one large central table and a set of smaller attendant tables displayed in a radial pattern around the central table. There is one large dominant table in the center of the schema. It is the only table in the schema with multiple joins (relationships) connecting it to other tables. The other tables all have only a single join attaching them to the central table.

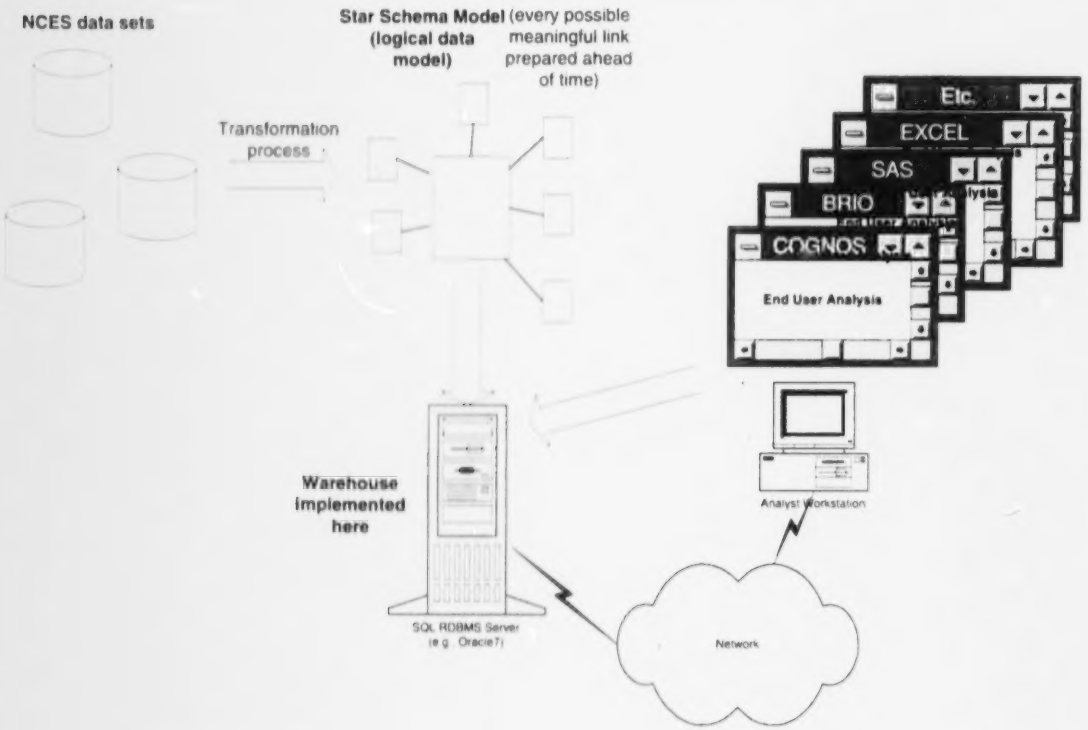
2. Implementation of the data model using Oracle, a standard relational database management system (RDBMS). A database management system consists of a collection of interrelated data and a set of programs to access that data. The collection of data is usually referred to as the database. A standard language for accessing relational databases is the standard query language (SQL).

3. Population of the model with data and metadata from multiple years of the CCD School District Universe Survey.

4. Demonstration of the model's utility with a commercial "front end" query tool called Cognos.

The conceptual model for the prototype system consists of components or processes that, taken together, meet the specific goals of the project. For each component or process, several alternative commercial products or methods may be available. The benefit of the standards-based approach is that any other component or process that supports the same common standards may replace any or all of the components or processes. To illustrate this approach, the data linking prototype can be viewed as consisting of components and processes as illustrated in Figure 5.1.

Figure 5.1
NCES DATA LINKING PROTOTYPE PROCESS



The following provides a description of each of these components and processes. These descriptions are of necessity somewhat technical.

1. A work process extracts data from relevant sources (i.e., NCES survey data sets) and loads it into a repository that is based on a standard "star-schema" or "dimensional" (the terms are synonymous) data model. This process can be accomplished via a wide variety of tools, methods, and technologies.

2. The star-schema repository is implemented under a standard SQL accessible RDBMS. SQL is a formal standard, controlled by ANSI (American National Standards Institute) and

BEST COPY AVAILABLE

endorsed via the National Institute for Standards and Technology (NIST) of the Department of Commerce as a FIPS (Federal Information Processing Standard). Examples of SQL RDBMS include Oracle7, Microsoft SQL Server, Sybase, Informix, and DB2. A star-schema repository can be quickly and easily transferred from one SQL RDBMS to any other. There are no practical, standards-based alternatives to SQL RDBMS for database management. There are certain data management products that are optimized to support star-schema data "warehouses" (e.g., Red Brick, Essbase). These products purportedly offer better performance than standard SQL RDBMSs in support of analytical queries against a star-schema database. The use of such products was beyond the scope of this task. In any event, the star schema model is equally as portable to this class of product, so no flexibility was sacrificed.

3. Data that are implemented as a star-schema under a SQL RDBMS can be accessed via a wide variety of end-user analytical tools (e.g. Cognos, Brio, SAS, Excel, etc.). These tools generate SQL queries that are transmitted to the RDBMS over a network. The RDBMS executes the SQL, and returns the results (data) to the end-user tool. These tools use one (or both) of two methods for communicating with the RDBMS. One method is through a mature, reliable *de jure* standard technology known as ODBC (Open Data Base Connectivity). ODBC allows any client tool that supports ODBC to communicate with any RDBMS that supports ODBC. ODBC is currently supported by every leading RDBMS product (including all of those identified above), as well as by every leading vendor of data warehousing analytical tools. (The second method is to use "native" data access drivers, i.e., to communicate directly with the RDBMS (using SQL), to avoid the processing overhead of ODBC translations. "Native" data access generally offers better performance than ODBC data access, however these differences are largely inconsequential for data access for analytical purposes). The goal was to identify leading products that can best illustrate the feasibility and benefits of linking survey data through a multidimensional or star-schema data warehouse model.

In summary, the data linking prototype may be considered as consisting of a set of components, where each component can be replaced with any other of the many that support the same basic technical standards. This mainstream standards-based approach was chosen so that the work products of this project will have the broadest utility and applicability toward larger NCES and Department of Education plans and goals.

5.1. ORACLE Warehouse

The first step in converting the SAS datasets into the ORACLE data warehouse was to create a crosswalk of the variable names and format types across all years of the CCD datasets. This crosswalk allowed us to change variable names and types so that for every year of data the same information had the same variable name and same data type.

Related to this standardization process was the standardization of the record layout. The variables on the datasets were in different order, depending on year. An identical record layout was made for each year to reduce the burden necessary when reading the ASCII files into ORACLE.

The percent of students in each ethnic group (white, black, Asian, Hispanic, Native American) for each district had been calculated and added in the CCD data sets. Summing up percents across districts would not make much sense, so these percents had to be backed out into raw numbers. This process also had to be performed on the Percent of Children in Poverty. The front-end tool will allow for the calculation of all percents. For a warehouse, it is important that all values be numerical.

The final step in SAS, prior to converting the datasets to ASCII, was to standardize some values in particular variables. For numeric variables such as the ethnicity variables, values of 'M', 'N', or '**' have been converted to null. This is NCES's way of indicating that data are Missing or N/A.

At this point data were converted from SAS data sets into ASCII files. These ASCII files were read into Oracle as a rectangular database. The rectangular database served as the starting point for the fact table. We then split off the dimension tables from the fact table in Oracle.

5.2. Cognos Tools

After a search of a number of commercial off-the-shelf vendors of OLAP tools, the Cognos suite of tools was chosen. Cognos was selected for its quality product, as well as their willingness to provide unlimited technical support during the creation of a prototype warehouse.

The two main tools of the Cognos suite are PowerPlay and Impromptu. The main tool, PowerPlay, has two views - Explorer and Reporter. In PowerPlay, a three-dimensional view of the data, called a PowerCube, is generated from a subset of the data available on the warehouse. An MS-Access database was used to create a PowerCube with NCES data.

In PowerPlay Explorer, the three-dimensional view of the data, or PowerCube, can be manipulated to display any of the included categorical data as rows, columns, layers (for 3 or more dimensional tables), or filters. Any one of the continuous variables included in the 3-dimensional view are used as the data in the main body of the tables (cells). This tool allows the user to drill up or down to any level of specificity pre-defined by the categorical data. PowerPlay is a very powerful exploratory data analysis tool. It allows for the quick calculation of the viewed data as percents, raw numbers, or even as any of a number of graphs or charts. See Appendix C for sample screens showing PowerPlay Explorer output for a two-dimensional table, a three-dimensional table, and a pie chart.

PowerPlay Reporter opens with a blank screen and from there the user specifies the exact variables wanted for rows, columns, layers, and filters. Here there are no restrictions on the types of variables used for rows or columns or layers as there are in PowerPlay Explorer. With this tool, tables that took many runs to create for table production are quickly created on the fly by the researcher with no programming at all. See Appendix C for sample screens showing PowerPlay Reporter output for a two-dimensional and a three-dimensional table.

Impromptu, the other major Cognos tool, is a report writer tool. This tool allows access to the entire warehouse of data. As it is linked directly to the warehouse, data access time is much longer than for either of the PowerPlay tools. The report tools available in Impromptu are much more limited than those available in PowerPlay. There is a simple one-way frequency, as well as a simple crosstab. The power of Impromptu lies in its ability to access the full warehouse. In PowerPlay the user is limited to a predefined set of variables. If a particular variable for an analysis is missing from any of the PowerCubes, then the researcher needs to go back to Impromptu to conduct analysis with that tool.

6. Lessons Learned and Possible Next Steps

The most important part of the NCES prototype data warehouse is its scalability and portability. Any of a number of parts of the warehouse implementation can easily be built upon. The data model can be expanded to include other dimensions relevant to sample survey data, such as sample design, imputation, and weighting dimensions. Since ORACLE was used for the database management system, any of a number of SQL-based relational database management systems can be substituted for ORACLE. The front-end tool, Cognos, can also be substituted with any other ODBC driven OLAP tool. The prototype shows the ability of a data warehouse to quickly deliver data in a number of user-specified formats.

A key to data warehouse flexibility is the use of a star schema model. The star schema model allows for any physical implementation and is flexible enough to handle operating system and software changes. The data warehouse tools industry has progressed to the point where they realize that each individual tool is only a small part of the overall warehouse.

Possible next steps include expanding the data in the dimensional model to include other NCES surveys including some sample surveys. This extension may require additional dimension tables to reflect sampling dimensions. A metadata dictionary that grouped variables across surveys into subject matter categories could also be added as a dimension into the data warehouse. A more comprehensive warehouse should also be evaluated by the user community for input on user-friendliness of the access tools before final investments in data warehouse software are made.

7. References

Ballard, Chuck. Strategies to Make Your Data Warehouse a Success. *TDWI Lesson From The Experts.* [On-line]. http://www.dw-institute.com/lessons/strateg.htm

Bernardi, John. Data Warehousing's Hidden Cornerstone: Information Storage and Retrieval. *TDWI Lesson From the Experts.* [On-line]. http://www.dw-institute.com/lessons/infostor.htm

Bohn, Kathy (1997). Converting Data for Warehouses: Understanding the Complexities and Resource Requirements Involved in a Quality Data Conversion. *DBMS Online.* [On-line]. http://www.dbmsmag.com/9706d15.html

Carlson, Bob (1997, March 24). *Secrets of Success.* [On-line]. http://www.computer world.com/search/AT-html/9703/970324SL12tekcol.html

Costa, Thierry. Three Issues in Data Warehousing. *TDWI Lesson From The Experts.* [On-line]. http://www.dw-institute.com/lessons/3issues.htm

Greenfield, Larry (1997). *The Data Warehousing Information Center Webpage.* [On-line]. http://pwp.starnetinc.com/larryg/index.html

The Data Warehousing Institute. *Ten Mistakes to Avoid.* [On-line]. http://www.dw-institute.com/papers/10mistks.htm

Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation (1997, June 19). Supporting State Efforts to Link Administrative Data Systems for the Purpose of Studying the Effects of Welfare Reform on Other State and Federal Public Assistance Programs (request for grant applications). *The Federal Register*, 62 (118), 33411-33417. [On-line] http://aspe.os.dhhs.gov/hsp/isp/frdata.htm

Demarest, Marc (1994). *Improving Data Legibility in Decision Support Systems.* [On-line]. http://vista.hevanet.com/demarest/marc/schema.html, a version of this paper appeared in *DBMS Magazine*, 7 (5), 55, in May 1994.

Drost, Karen. Lessons Learned: Three-tiered Relational OLAP. *TDWI Lesson From The Experts.* [On-line]. http://www.dw-institute.com /lessons/3tier.htm

Edelstein, Herb (1997, April 21). *Mining for Gold.* [On-line]. http://techweb.cmp.com /iw/627/27olmin.htm

Elkholy, Marianne. Great Performance in Date Warehousing. *TDWI Lesson From The Experts.* [On-line]. http://www.dw-institute.com/lessons/grtperf.htm

Feldman, Steve. Experts Tell Why Data Re-engineering is Required for Data Warehousing. *TDWI Lesson From The Experts.* [On-line]. http://www.dw-institute.com/lessons.reeng.htm

Finkelstein, Robert. *Understanding the Need for On-Line Analytical Servers.*

Glassey-Edholm, Katherine. Keys to the Data Warehouse: The Role of query Tools in Successful Date Warehouses. *TDWI Lesson From The Experts.* [On-line]. http://www.dw-institute.com/lessons/rolequer.htm

Hill, Neal. The Warehouse is Not the "Answer". *TDWI Lesson From The Experts.* [On-line] http://www.dw-institute.com/lessons/not_answ.htm

Horrocks, Chris (1995). *Data Warehouses: An Executive Information Perspective.* [On-line]. http://www.csc.com/about/tech_dw_arch.html

Houde, Lisa, & Saylor, Michael J. True Relational OLAP. *TDWI Lesson From the Experts.* [On-line]. http://www.dw-institute.com/lessons/rel_olap.htm

Imirie, Peggy. Your Data Warehouse: A Business Success or Science Project? *TDWI Lesson From The Experts.* [On-line]. http://www.dw-institute.com/lessons/sciproj.htm

Kimball, Ralph (1996, August). Dangerous Preconceptions: Discovering the Liberating Truths That Can Lead to a Successful Data Warehouse Project. *DBMS Online.* [On-line]. http://www.dbmsmag.com/9608d05.html

Kimball, Ralph (1996). The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses. John Wiley & Sons, Inc.

Kinikin, Erin. Tuning Techniques for Interactive Data Warehousing. *TDWI Lesson From The Experts.* [On-line]. http://www.dw-institute.com/lessons/tuning.htm

McElreath, Jack (1995). *Data Warehouses: An Architectural Perspective.* [On-line]. http://www.csc.com/about /tech_dw_arch.html

No author. (1996, February). Meta Group's Karen Rubenstrunk: Standardizing Metadata. *DBMS Online DBMS Interview.* [On-line]. http://www.dbmsmag.com/int9602.html

The OLAP Council (1995, January). *OLAP and OLAP Server Definitions.* [On-line]. http://www.arborsoft.com/olap/terms.html

Porter, Patrick L., & Radcliff, Deborah (moderators) (1997, June). CIO Roundtable: Date Warehousing for Grown-ups. *Software Magazine On-Line.* [On-line]. http://www .sentrytech.com/sm067dw.htm

Raden, Neil (1996). Modeling the Data Warehouse. [On-line]. http://members.aol.com /nraden/iw0196_1.htm

Rist, Richard A. Is Data Warehousing Advancing Your Thinking? The Parallel Revolution. *TDWI Lesson From The Experts.* [On-line]. http://www.dw-institute.com/lessons/parallel.htm

Sachdeva, Satya (1995, December). *Metadata: Guiding Users Through Disparate Data Layers.* [On-line]. http://www.inquiry.com/publication...ec95/fe1205.ADT19951201FE 1205.html

Sokol, Marc. The Next Generation of Data Warehousing. *TDWI Lesson From The Experts.* [On-line]. http://www.dw-institute.com/lessons/nextgen.htm

Taplin, Caroline (1997, April). *Abstract of Project: State Linked Data Meeting and Monograph.* [On-line]. http://www.os.dhhs.gov/cgi-bin/waisgate?WAISdocID=55886 25542+3+ 0+0&WAISaction=retrieve

Yost, Mickey B. Using the Star Schema to Access the Historical Data of the National Agricultural Statistics Service. National Agricultural Statistics Service, U.S. Department of Agriculture, Washington, DC

Welch, J.D. Business Requirements Analysis: The Missing Link of Data Warehousing. *TDWI Lesson From The Experts.* [On-line]. http://www.dw-institute.com/lessons/req anal.htm

Wood, Edward E. Jr. Middleware: The Foundation for Data Warehousing. *TDWI Lesson From The Experts.* [On-line]. http://www.dw-institute.com/lessons/middle wr.htm

Zimmer, Harry. Data Warehousing: Are You on a Path to Success or Failure? *TDWI Lesson From the Experts.* [On-line]. http://www.dw-institute.com/lessons/rghtpath.htm

APPENDIX A

SUMMARY OF INTERVIEWS

Tim Newell, Sierra, (703) 522-8628, ext. 202

Has linked the CCD agency file to the Census F-33 file, but has only within years and not across years. Said that Pinkerton has worked on linking IPEDS. Pinkerton took five IPEDS survey files within a year and linked them. Pinkerton has also worked with linking SASS across years. Suggested contacting Marge Sterner (Pinkerton, 703-820-5571).

Tom Snyder, NCES, (202) 219-1689

He has linked multiple years with many different databases. Most of his linking was not at the record level; he aggregates data within years and then compares across years. Said his linking work has been ad-hoc and that if he has any documentation it is in the form of program code.

Mentioned there is an IPEDS file with 15 years of linked data for Doctorates and the data may be located at Pinkerton or NSF. He was not very hopeful because the person who created it has since retired. The biggest problem with this file was caused by schools changing their IDs.

Currently working on a school district universe database with 10 years of data. He is matching the Census F-33 file to CCC District data (nonfiscal). Encountering problems with the older files having more mismatches. The documentation for this file is pretty far along and can be picked up.

Also has a Lotus file which contains state level CCD enrollment by grade for 1965-92. Most of the merges are for one-time uses and he does not keep the data sets created. He does keep the programs in case there is a need to recreate the data set.

Noted NSF has a fairly extensive linked data set for IPEDS called CASPER which contains most of the major variables and it is on CD-ROM.

Reported that Nabeel Alsalam at NCES is working on linking NELS and IPEDS and that he also has done something with 10-15 years of CPS data, the results of which he has on CD-ROM.

Steve Owens, Governments Division, Census, (301) 457-1586

Working on the integrated sampling project. Has done some work linking SASS, CCD, and QED. Found linking SASS LEA and CCD district easy, but had some problems linking SASS and CCD at the school level because the school IDs were not matching (Comment: this could be because he is using the public SASS data which recoded and scrambled the school IDs). Had problems linking the 1987 SASS to CCD because the 1987 SASS uses APIN which is not a CCD ID number. Thinks there must be a crosswalk for APIN to CCD ID somewhere but he that does not know where.

Had two small findings. (1) When he linked CCD and IPEDS discovered that some schools are on both files. This is not documented anywhere. This occurs because some higher education

institutions around the country are becoming more involved in helping to run school districts. This type of occurrence showed up most often in Colorado, Indiana, California, and Michigan. (2) Found there were some schools on both PSS and CCD. This may be caused by the use of school vouchers or public school districts contracting out work to private schools.

Has also done some linking of the 1990-91 SASS and the 1992-93 CCD.

Stated he does not have any products yet but that a draft report is available from Jonaki Bose (NCES).

Bill Freund, NCES, (202) 219-1373

Had a great deal of information on current and previous NCES data linking projects and many people to contact. Said IPEDS has been linked to all postsecondary surveys since IPEDS is the universe. Provided the following people to contact about the existence of this documentation and any linked files which may exist.

NPSAS: Drew Malizio, NCES, ext. 1448
NSOPF: Linda Zimbler, NCES, ext. 1834
B&B: Paula Knepper, NCES, ext. 1914
Recent College Graduates: Peter Stowe, NCES, ext. 2099
HS&B: Dennis Carroll, NCES, ext. 1774

Mentioned that Dennis Carroll knows the most about linking of all these surveys.
Said that Sam Barbett linked IPEDS with the Office of Postsecondary Files (these files consist of institutions eligible for student aid). Noted Census is doing some work (this is the integrated sampling scheme work) and that the F-33 (a Census file) and Census mapping files have been linked to the SDDB.

Arnold Reznek, Center for Economic Studies, Census, (301) 457-1856, areznek@census.gov

The Center for Economic Studies has linked CCD across years from 1987-94 and has also linked SASS back to CCD. The purpose of their project is to link SASS back to CCD. They are in the process of evaluating problems with linking the CCD data across years. They expect to have a draft report on this by June. What they are doing is very much a pilot. They do not have any data sets but they create data sets using SAS and then delete them when they are done because they are so large. They have linked schools and LEAs (districts) over time.

When asked if they have anything the NCES project could use now he said that they do not; they only have the programs.

Some areas they have been exploring are having school IDs changed, and if there are any broken links. They are trying to figure out what questions people may try to answer using linked data and what links are useful for analysis. Some of the questions under consideration are: has a school

changed its physical location or has a school site (location) had many different uses (what kinds of uses a facility has had over its life)? Suggested there should be an ID on the file which identifies the place, the physical location, of a school building.

They have also looked at measuring school quality using information from SASS and CCD. They have also looked at what happens to the SASS sampled schools in the period between when they were drawn for the sample and when they are surveyed. They are doing this by looking at the CCD data for the in-between years.

For software, they have purchased an SQL server only because this is what NCES uses. Added they are not database package experts.

Katherine Chandler, NCES, (202) 219-1767

For NHES they commonly link the survey files with a year. There is no longitudinal component so linking files across years cannot be done and is meaningless. The files which have been linked within years are:

1991: Adult Education and Early Childhood Participation have been linked when interview came out of the same household.

1993: School Readiness has been linked with School Safety, Discipline Parent, and Discipline Youth. The Discipline Parent and the Discipline Youth files have also been linked.

1995: Repeat of 1991, Adult Education and Early Childhood were linked.

1996: Parental Involvement has been linked with Household Involvement. There are four files which can be linked:

Parental Involvement
Youth Involvement
Adult Involvement (This can be linked with Household only)
Household Involvement

The data have been analyzed by looking at the differences of parent and child perceptions, differences of civic and parent involvement with a child, and how adult education differs for those people with and without kids.

There is documentation on the linking which has been done and it is in codebooks and manuals. It has been necessary to redefine and rename variables so they do not get overwritten on a linked data set.

It is possible to link on Zip code for the restricted file and some of this has been done by linking to Census data (STF3B file). Also knows of linking the National Crime Victimization Survey School Crime Supplement with CCD and PSS school information.

Thought Data Analysis Systems (DASs) were not appropriate or useful for NHES.

Bill Fowler, NCES, (202) 219-1921

Has a data set which linked students from three waves of NELS. (Did not use dropouts, the student was in the same school for all three waves, and the student had complete reading and math scores for all three waves.) to CCD school district that were not missing expenditures. Does have longitudinal weights on the file, these are the panel weights from NELS.

Said the files merged pretty well. Looking at achievement and expenditures, and is controlling for socio-economic background, test scores, and school size. He will be reporting his findings to AERA.

There will be a forthcoming report on the findings. Working on this report with Jeff Owings. Has the SAS programs that he used to create the linked file, but does not have any documentation on resolving definitional differences. The data set created is restricted. Believes the data will lose its usefulness if it is transformed into a public use file. Mentioned NELS is school and not district representative.

Said there are a lot of people interested in linking school data to finance data. They use this to try and determine if money matters to students.

Identified Nakib Yasser (302-831-4227, NAKIB@udel.edu), an assistant professor at the University of Delaware, who has linked SASS to CCD financial data.

Drew Malizio, NCES, (202) 219-1448

IPEDS is the universe file for NPSAS. Said the students are of interest (which are the NPSAS data). They do conduct longitudinal studies with NPSAS and they do have weights.
Said when they encounter problems linking it is most often caused by institutional characteristics changing. Noted IPEDS has locale codes so it is possible to match student Zip codes to locale codes and then to Census data, but they have not tried to do this.

The types of questions answered using this linked data include: the types of institutions people are attending, the race of students who are attending certain types of institutions, the percentage of students attending institutions requiring the ACT and/or the SAT.

They have no formal documentation. They have encountered some problems with linking that data because the way the sample was selected has changed. Questions on the survey have also changed because much survey work is done by phone now.

Sam Barbett, NCES, (202) 219-1592

Has done linking with IPEDS files across years. Most linking is add hoc and it is not formally documented. They link on unit ID which does not change. Helped Linda Zimbler (NCES) link some IPEDS data to NSOPF. They do encounter some problems when schools merge, for when schools *disappear* it is not always due to a closure, but to schools merging.

They have not linked to non-NCES data but that they are in the process of adding an ID to allow linking to the financial files. There is a code on the IPEDS files which allows them to be linked to the HEGIS surveys done in 1967-1985. HEGIS is the predecessor to IPEDS. Most of the documentation is add hoc and in programs.

Frank Johnson, NCES, (202) 219-1618

Has not done any linking of CCD across multiple years. Linking has been done within years and linking has also been done to the Census F33 file. Within years the CCD school and LEA universe files can be linked. The first seven characters of the school ID are the same as the seven character LEA ID. He said that the majority of the work he does is quality checking of what Census does.

As to longitudinal data, he said that they are universe files and that they assume they are universe files even though they know that CCD does not contain or is missing some schools (e.g., schools on military bases are not included in the CCD, and he does not think they get all of the BIA schools).

All of the documentation exists in the documentation package for each survey. They expand the surveys, collecting information on more items, and they may take a previous survey item and break it out into several items.

An example of a problem they encounter is when a teacher is not picked up at the school level but shows up at the LEA level. This happens when the teacher is hired by the LEA and works at multiple schools.

Believes it is appropriate to analyze CCD data linked both within and across years.

John Sietsema, NCES, (202) 219-1335

Said AIR has linked CCD across multiple years and the best contact at AIR is Don McLaughlin. Sietsema said that AIR did an analysis on all available CCD data on the CD-ROM. Thinks they used the years 1986 to 1992. What AIR did was run quality and consistency checks on the data. They looked for records that were in one year and not in another. If they found a case where a school was there one year, gone the next, and then back the next year, they tried to fill in for the missing year in the middle.

AIR was linking the data so they could write a report, "Trends in School District Demographics." For this report AIR only used schools and districts that had data for each CCD year.

Suggested talking with Valerie Conley (formerly Synectics) because at Pinkerton she developed a file/program which enabled people to identify a record on CCD and then find its past history.

Said some of the problems people encounter for CCD were that some states changed local code when they were not supposed to. In CCD once a record is removed its ID should be removed and not used again: if someone goes out of CCD and then comes back in, they need a new ID on re-entry.

Said the data do still exist, but thinks it may be better to start from scratch. When AIR did this they did it as they went along, so the documentation may not be very good.

Recommended when working with CCD to get very familiar with the data. Can check with him if there are any questions. Examples of questions that people overlook or misinterpret include deleting a district record if its membership is zero. In such a case children could be assigned to another school (a home school) or students sent to a vocational school in that district. There are some districts which do not operate schools but pay tuition to other districts, so those students show up in the other districts numbers.

Kerry Gruber, NCES, (202) 219-1461

Provided an overview of what SASS data have been linked to and what people have attempted to link SASS data to. She said that Larry Picus (USC, 213-740-2175, Assistant Professor, National Center for Finance and Productivity in USC's Department of Education) linked the 1987-88 SASS to CCD and then to the Census F-33 file. Picus linked SASS school to SASS district. He then linked this to CCD agency (district). And then finally he linked this to Census F-33 district finance data.

Said Mike Podgursky (University of Missouri, 314-882-4574, Chair of the Department of Economics) may also have linked SASS data as he is interested in labor market conditions for teachers. Thinks he may have some occupational data that he links with SASS.

Mike Podgursky, University of Missouri, (573) 882-4574

Has linked SASS with measures of costs of living from CCD. Has linked SASS with the USDA Beal Code a 1-to-10 scale which describes how rural an area is. Has used the 1987-88 and the 1993-94 SASS to judge the quality of teachers by categorizing the institutions they graduated from using Barrons codes. There was an IPEDS institution code on this file. Has also linked Census data with SASS to compare teacher with non-teacher salaries in the area. Has used median home price from Census to judge the quality of an area.

Questions he and other researchers were trying to answer were how teachers and nonteachers compare. They try and find relative teacher pay compared to cost of living and other jobs in the area.

NELS is used more than SASS because NELS has outcomes variables while SASS does not. SASS is an ideal file for linking (mapping), but it needs to have outcome measures put on it.

He has the data and has worked largely with the end product. Suggested contacting his other researchers to find out the specifics about the data and documentation.

Said state level achievement tests could be put on, or fringe benefit information, or possibly specific data on alternate jobs from BLS. Thinks SASS should be used to fill in the gaps for what the states fail to collect, and that SASS should focus on what drives teachers' and administrators' decisions and asset allocation. He said that nowhere does anyone ever ask a school how many teachers did not get their contracts renewed because they were bad teachers; did teachers turnover (leave or were fired) because they left for greener pastures or they were bad teachers.

Bill Sonnenberg, NCES, (202) 219-1580

Has linked CCD data with Census (F33 financial data) for the past 20 years. Has also done some work linking Census data with school data to determine Title I allocations. When linking this data used school district ID. Has not done any longitudinal analysis with this data and said the main problem people would have looking at this data over time is that no one has kept track of changes in school district IDs over time.

The main problem encountered in attempting to link data is that the IDs do not match. There is no documentation for his linking, all his linking was done Ad Hoc. The linking was done in SAS. He does not *comment* his programs and he needs to spend several hours looking at them to figure out why he wrote the code he did. He said that the only time the data was linked and documented was back in the 1970s for one study.

This linked data is used to analyze resources with other information such as teachers and instructional expenditures or Title I poverty by school district. Mentioned that court cases are using CVs more and more as a measure of disparity.

Dale Ballou, University of Massachusetts, (h) (860) 298-9964

Did not remember much of anything of the linking they have done, but did have some good ideas for improving NCES surveys and some things that our analysis system should be able to do.

He said that in general NCES surveys need to have better geographic linking capabilities (e.g., Zip code, area code). Has worked with follow-up surveys and that they have nongeographic information on teachers. There is no way to tell what labor market they left and what labor market they went to.

Said in order to analyze labor market data you need to have strong geographic identifiers. Should make sure that what is developed has strong geographic data linking capabilities.

Said it is very easy to find data to link in when the data are aggregated to the county and state levels.

Peter Stowe, NCES, (202) 219-2099

Has not linked any NCES data in four to five years when he linked IPEDS with Recent College Graduates (since been replaced by B&B). Did this on his own and it was not done to create any final product. Said it was done Ad Hoc, was very low quality, and not designed for others to use.

Said what project will attempt to do is important for analysts. Offered some potential data sources and areas for links. Said would be beneficial to link NCES data with BLS data in particular labor market outcomes. Over the past five to ten years people have turned more to labor market outcomes. People are looking more at the education outcomes and how they translate into labor market outcomes. It used to be enough to state number of degrees earned, now want to know how and which degrees translate into jobs. Before surveys would report that X percent of the population received BAs now to know that X percent of the population with BAs got what jobs.

Larry Picus, USC, (213) 740-2175

Back in 1990 linked the 1987-88 SASS to the Census F33 (Census of governments data). Said it took them over a year to do and they had to hand check all of their links. Said that after this experience he would never try to do anything of this kind again. He had graduate students do the programming and that the quality of the file is about medium. He also said that the file may still exist on their UNIX system. He was using these data to gauge how education resources are allocated and used.

Would like NCES and Census data in which a district can be easily matched from one file to another. Would like a relational database which he could access (on-line) to create and download linked data for analysis.

Yasser Nakib, University of Delaware, (302) 831-4227

Has a very small and specific high quality data set that he has not shared with anyone. He used CCD as an intermediary to link NELS student level outcomes with data which states collect. He has done this for Florida. Basically NAEP, SASS, and NELS all have the NCES ID, while data from the states have a state ID. Used CCD which has both the NCES and state ID to link NELS with the state data. Goes to the states to get the data.

Has no formal documentation on his programs which are all done in SPSS. Said it was hard at first but that now it is easy with practice. Would like to see a good connection established between NCES data and state data. Said states can produce much more detailed and useful information than the federal government can ever collect.

Don McLaughlin, AIR, (415) 493-3550

Has linked CCD from 1986-87 to 1993-94 for what he classified as regular public school districts only (he did not use all of the CCD data). Said it is very easy to get a 95% to 99% link on ID. The file has been cleaned up to eliminate annoying anomalies and inconsistencies, such as student enrollments going from 100 to 1000 then back to 100.

On this linked CCD file he did a fair amount of longitudinal editing and imputation (some of this was done using PROC IMPUTE, a FORTRAN program he wrote). Said an ESSI contract has been has proposed to develop a longitudinal CCD system, so one would only need to add the new data and the run a program which would clean the data and make the file consistent. Said NCES was waiting on awarding this until the prototype analysis system is done.

Has encountered the typical problems when working with the CCD data, including the closings and mergers of districts. For the future, has concerns that CCD will not collect data on Charter or home schools.

Said there is an appendix to a forthcoming report which describes his imputation procedures. If a copy is needed, it can be obtained from Lee Hoffman (NCES). Said the quality of the data is higher than that of the CCD data. Also said he would be willing provide a copy of the data. Thinks it would take about $50,000 to complete the cleaning on this file.

Has also been experimenting with linking CCD, SASS, NAEP, and State Assessment data for 19 states. The State Assessment data is collected by the states and is not standardized, a standard form is not used.

Don McLaughlin was contacted for a copy of his linked CCD data and the programs. There should not be a problem with getting a copy of the data, but there may be some with getting copies of the programs. Said he did not want the programs reviewed by other people and judged saying what he did correctly and incorrectly. He said that in his programs there are sections of code he did not run and it is not apparent what code was run from looking at the programs.

Don McLaughlin asked for an e-mail request on materials. Said he has a proposal in with NCES to complete editing the data. NCES is aware that editing still needs to be completed on his data but that NCES would like us to look at his data. Likely will need to set up a billing vehicle especially to use his data to develop the prototype.

Jay Chambers, AIR, (415) 493-3550 ext. 8111

Linked SASS teachers and administrators to SASS schools then to SASS LEAs then to CCD district then to Census county level data (from the City County Data Book). Did this for the 1987-88, 1990-91, and 1993-94 SASS surveys. He then when back and matched districts across years for CCD. The only problem encountered was that some SASS teachers did not have an LEA ID, because of LEA nonresponse. He said that he had to have someone work on a crosswalk for the three rounds of SASS to find consistent variables across all three years.

Said that when they did analysis on the file they did it within year and used the SASS teacher weight. When he analyzed the data across years he did not use any of the SASS data so it was not necessary to use weights.

As for documentation he said that there will be a technical report released by the end of April which will have a general description of what they did in an appendix.

Did their programming in SAS. The data still exist and a copy can be obtained if necessary.

Talked about a few other painful linkings they have done. Has a file which links each central city in the US to its nearest district. They went through and computed the distances of every central city from each district in the US.

Found the distance of each district from one of the 280 or so weather stations in the US. They did this so that they could put on climate information and see if there is a relationship between climate and teacher salary. They also used data from the FBI uniform crime reports and matched the names of the city of the crime report to the name of the CCD school. This enabled them to attack crime data to the schools. What made the linkings so painful was that they almost had to match a lot of the city names by hand. The city names on the FBI data may not exactly match the city names on the CCD data.

Valerie Conley, formerly Synectics

NSOPF has been linked to IPEDS by using the institution ID which is the first six digits of the nine digit faculty ID. They have also linked NSOPF faculty and institution data. Valerie also said that they have a cross walk file for NSOPF ID to IPEDS. She said that the NSOPF IPEDS linked data for 1992-93 is on CD-ROM, and possibly that the 1987-88 linked data may still be on the mainframe.

Steve Wenck of Synectics did the link between NSOPF and IPEDS (when he was at Pinkerton) and that the code still exists.

Worked on a data product called CCD Net which was a prototype system to link CCD data both within and across years from 1986 to 1992. This system can link from schools to districts to states both within and across years. There is a users manual but that there is not documentation on the development of the executable.

Said that when they were doing these linkings that the main problem they had were inconsistencies, not many so the matches and links worked very well. Said they developed rules for how to resolve inconsistencies but there is no documentation for this they just wrote the code needed to resolve the inconsistencies.

Mentioned that Speedy Express is also a data linking type of system (on NCES's web site in "edi").

Paula Knepper, NCES, (202) 219-1914

Has linked IPEDS with BPS, B&B, and NPSAS. Did this mostly to get institutional characteristics onto student records. Has linked both within and across years. She said that the majority of the files she deals with are universe files so that there are no weighting issues; in the few cases where there have been weighting issues she was able to use the weights from the student file (IPEDS).

The only major problem she encountered was institutions changing their ID's across years. She has not encountered may problems caused by definitions changing across years. She said that she has been lucky because most of her definitions have not change across years. She said that she has no formal documentation most of it was done add hoc. She has linked her data using FORTRAN, SPSS, SAS, COBOL, and Pascal. Some of her linked data does still exist and she said that some of it is good and some of it is not.

She said that the number of analyses she has done with the linked data are too numerous for her to list/remember. Most of the time she needs to link data one way to answer a question and then link data another way to answer another question. So there is not much continuity in the analyses that she is doing on the linked data. Most recently she has linked B&B with IPEDS and NPSAS with IPEDS. She did these links to answer questions about graduate tuition and enrollment.

Mickey Yost, USDA National Agricultural Statistical Service (202) 720-3649

Mickey Yost, the NASS Data Warehouse Manager, said that USDA does over 800 surveys a year comprising over 10,000 survey items. USDA uses Red Brick for their data warehouse and then they use Brio and SAS as front end tools to go against it. He said that the way they have set up their system using Red Brick that it is impossible for users to make illogical queries. From my discussion with him it sound like USDA is doing exactly what NCES is moving towards doing. Mickey Yost invited us down to take a close look at what they are doing, I said that we may take him up on it. He also said that when developing a system like this, one that allows querying of survey data, it is important to know some of the ways that people will look at the data. He said that all of their data in SAS data sets. He added that an insignificant amount of their data is in Lotus spreadsheets, but it is a tiny, tiny amount.

Fan Zhang, Synectics (703) 807-2306

Linked 1990-91 SASS Public School file with both the 1989-90 and 1990-91 CCD School Universe files at the school level. The linkage between the SASS and CCD files was through an identification variable called "CCDIDSCH". The purpose of the linkage was to adjust SASS state estimates (in 40 states) of the number of teachers to be more consistent with the CCD estimates which were considered to be more accurate. SASS estimates were known to be at least 15% higher than CCD estimates of teachers in ten states because the school administrators did not report schools in the same way in the 1990-91 SASS as in the 1990-91 CCD. For example, a school with grades K-8 at one address might be two CCD schools - an elementary school with grades K-6 and a middle school with grades 7 and 8 resulting in SASS reporting more grades than the same school had on the CCD.

A Statistical Analysis System (SAS) program was written to compare the CCD and SASS datasets and to adjust and modify the 1990-91 SASS Public School data file. A copy of the program was included as Appendix B of the final report (F. Zhang, M. Saba, and B. Scanlon, *CCD Adjustment to SASS*, July 1994). Since the sampling frame for 1990-91 SASS was the 1988-89 CCD, 155 schools in the 1990-91 SASS sample were not found in the 1990-91 CCD. When the adjustment criteria was applied to these schools, it was concluded that 29 of them needed to be adjusted. It was necessary to then match these schools to the 1988-89 CCD to obtain the data necessary for the adjustment. There were a total of 300 schools in SASS that were adjusted. Estimates were generated for five characteristics of schools based upon a set of rules: the number of full-time equivalent (FTE) teachers, total student enrollment, the number of Hispanic students (grades K-12), the number of students participating in extended day or before- or after-school day care, and the number of students who receive free or reduced price lunches.

Sameena Salvucci, Synectics (703) 807-2309

Selected estimates from the CCD School Universe, LEA Universe, and State Nonfiscal files were compared across three years, 1988-89, 1989-90, and 1990-91 in order to measure the internal accuracy of the CCD. The following counts were compared:

- *Aggregated Student Counts*: Comparisons were made for each state and the District of Columbia and the U.S. Territories for each of the three years between the following pairs of CCD surveys:

    1. The School Universe and the Local Education Agency (LEA) Universe;
    2. The School Universe and the State Nonfiscal;
    3. The State Nonfiscal and the LEA Universe.

- *Aggregated Full-time Equivalent (FTE) Teacher Counts*: Comparisons were made for each state and the District of Columbia and the U.S. Territories for each of the three years between the following pairs of CCD surveys:

    1. The School Universe and the LEA Universe;

2. The School Universe and the State Nonfiscal.

Selected estimates from the 1990-91 CCD and 1990-91 SASS surveys were compared. Comparisons were made for each state and the District of Columbia and the U.S. Territories The following counts were compared:

- *Aggregated/Estimated Student Counts*: Comparisons were made between the following pairs of surveys:

    1. The SASS School Survey and the CCD School Universe;
    2. The SASS Teacher Demand and Shortage (TDS) Survey and the CCD LEA Universe;
    3. The SASS School Survey and the CCD State Nonfiscal Survey

- *Aggregated/Estimated Full-time Equivalent (FTE) Teacher Counts*: Comparisons were made between the following pairs of surveys:

    1. The SASS School Universe and the CCD School Universe;
    2. The SASS TDS Survey and the CCD LEA Universe;
    3. The SASS School Survey and the CCD State Nonfiscal Survey.

- *Number of Schools*: Comparisons were made between the SASS School Survey and the CCD School Universe survey.

- *Number of LEAs*: Comparisons were made between the SASS TDS Survey and LEA Universe survey.

Statistical Analysis System (SAS) programs were written to compare estimates within CCD and between CCD and SASS. These self-documenting programs are still available at Synectics. The programs used to compare between CCD and SASS included adjustments for definitional differences. The results of these comparisons are documented in detail in a final report (S. Salvucci, L. Thurgood, G. Carter, and R. Lerner, *An Examination of the Quality of the Common Core of Data*, 1998).

APPENDIX B

Longitudinal Editing and Imputation of CCD Data

The specific editing and imputation steps taken by McLaughlin are described in this section. Chronologically, the 1986-87 through 1991-92 data were edited and imputed simultaneously, and the 1992-93 and 199394 data were subsequently imputed using the values from the preceding years. The editing and imputation was performed in the following 15 steps.

Step 1. Specify the records to be included. Identify school districts that change type from regular to non-regular and back, and set the type to be constant. Reported types of some districts in Maine, Massachusetts, California, Ohio, Virginia, and Vermont were changed in some years. (For one LEA on the Mississippi River whose state did not match its identification code, the variable STATE was changed.) Also, if any district has no students, no teachers, and no schools, and does not merge with any schools on the school file, in any year, delete it from the **file.** This step determines the number of district records on each year's file.

Step 2. YEARS. Create YEARS, a string with one character for each year: "Y" if the district is on the district file and merges with at least one school on the school file in the year, "N" if the district is on the district file but merges with no schools on the school file in the year, and "M" if the district is not on the district file in the year.

Step 3. Number of schools. If the number of schools is missing for a district for a year, use the number from a preceding year with data. If the number is not available for any year, use the number of records on the school file for the district. (If none, set the number of schools to zero.)

Step 4. Grade span. If high grade and low grade are missing for a year, use the previous or closest year if some year has data. Otherwise, impute from school file. If the school file grade span is indeterminate, but there is a school, impute KG-to-12. Otherwise (if there is no school), impute as missing. Edit gradespan to remove cases in which low grade is higher than high grade-set them equal to whichever is not imputed, or if neither is, to the lower of the two.

Step 5. Number of teachers. Set spurious zeros for numbers of teachers (in Massachusetts and Michigan in 2 years) to missing. If number of teachers is missing in a district for a year, use the sum from the school file if there is a match. Otherwise, use a prior year's count, or if no teacher counts are available for any year, impute a value equal to the product of the number of schools times the number of grades in the gradespan (i.e., one teacher per school per grade). If the gradespan is indeterminate, impute one teacher per school.

Step 6. Edit number of students. Replace zero or missing values for enrollment in a district, or values that differ from an adjacent year by both 40 and 40 percent, with positive values from the school file whenever available. Note that when single years were added to the file later (i.e., 1992-93 and 1993-94), this step was repeated.

Step 7. Edit student/teacher ratio. Remove large or inconsistent student/teacher ratios (S/T). If for some year, a district's S/T is greater than 50 or S/T is "inconsistent" with both of the 2 adjacent years (by a factor of 2 or more), and the adjacent years are consistent with each other, then either set S to missing (to be imputed) or impute T directly. If S is consistent with adjacent years but T is not (each by a 40 percent factor), impute T as the average of the two years it is

adjacent to. Otherwise set S to missing. One district, new in 1991-92, has number of teachers imputed from 1992-93, because its number of teachers in 1991-1992 created a student teacher ratio greater than 700.

Step 8. Impute number of students. Run PROC IMPUTE to impute total students in the 6 years. The imputation is BY two categories of number of schools (districts with fewer than 4 schools and districts with 4 to 19 schools). No districts with more than 20 schools were missing total enrollment. The average number of schools and average number of teachers were used in PROC IMPUTE.

Step 9. Racial-ethnic percentages. This step imputes ethnic distributions. First, the SDDB (1990 Decennial Census, mapped onto school district boundaries) is used to obtain percentages of each district's child population in different ethnic groups. For 27 districts for which no ethnic data are available for any year on the CCD or for the SDDB, impute the average for districts in the same city, or if not available, from the same county. For districts with data in some years but not others, perform the edit check described below, then use PROC IMPUTE. (However, no ethnic data were available for 1986-87, and none were imputed. Ethnic distributions for that year are not included in the report.)

Set inconsistent values to missing. These are values for districts that have values for at least 3 different years, and at least one of the percents differs from the average of all years by both (a) at least 25 percentage points and (b) at least 5 standard deviations. Also, for convenience, set the percentages for districts with zero students to the national averages: 1.1, 6.1, 5.4, 2.2, 85.2, for Asian, black non-Hispanic, Hispanic, Native American, and white non-Hispanic, respectively. Run PROC IMPUTE with the 20 variables (four ethnic groups (excluding white non-Hispanics) for each year from 1987-88 through 1991-92). An additional run using all years' data, but only imputing the last 2 years, was made to impute missing values for 1992-93 and 1993-94.

If the resulting sum of the minority percents is greater than 100 for any district, they are normalized to 100. The white non-Hispanic percentage is set to 100 minus the sum of the other percentages in all districts.

Step 10. Locale code. For districts with schools with locale codes, the NCES standard procedure for deriving district locale codes from school locale codes was used. That procedure assigns the most frequent school locale code in the district, setting ties to the more urban local, with the possible exception that for districts in which at least three-fourths of the schools have locales spread among values of 1, 2, 3, or 4 (i.e., in metropolitan areas) but the most frequent single school locale is 5, 6, or 7 (i.e., large or small town or rural), the district locale would be set to the most frequent of the values 1, 2, 3, or 4. (That exception did not occur in these data.)

For districts with no locale code in any year, the most frequent locale code for districts in the same county was used. If no data were available for the county, (a) the value 2 was imputed if the metro status code was 1; otherwise, if the number of schools was less than 5, the value 7 was imputed. If the metro status code was 2 and there were 5 or more schools, the value 3 was imputed; and if the metro status code was 3 and there were 5 or more schools, the value 6 was imputed. These rules are based on minimizing the percent errors based on relations observed for

districts with data. Although the locale code was imputed separately by year, imputed values for a district were forced to be constant across years, equal either to the latest unimputed value or, if there were no unimputed values, to the modal value.

Step 11. Percent of school-aged children in poverty. (This variable was taken from the SDDB. It was therefore missing for all CCD districts not present in the SDDB.) The average percent poverty for districts in the same county was used to impute percent poverty. If there were no districts in a county with data, the average value 17 percent was used.

Step 12. Counts of special education students. First, counts in all districts in states which reported uniform zeroes in a year were set to missing, to be imputed. Second, if the number in a district exceeds the total number of students for a district, it was imputed to be equal to the total number of students.

Counts were then translated to fractions of total enrollment, and two variables were created-the average fractions for 1987-88 and 1988-89, and the average fractions for later years. Two averages were used because the values in the earlier years were not highly con-elated with the values in later years. PROC IMPUTE was run, with five special education percentages (one for each year from 1987-88 through 1991-92), the two overall averages, and the percent of enrollment that was black non-Hispanic, plus Native American, minus Asian. It was run with separate hot deck distributions depending on whether there was a determined gradespan. These variables were selected on the basis of regression model results. Imputed percentages were translated back into counts.

Step 13. Four types of high school completers. Data were only available for the years after 1986-87, and the high school equivalence results were not available for 1991-92. First, values for 12th grade enrollment were imputed (and later dropped), in order to impute graduates as a ratio to the preceding year's 12th graders. Imputation of 12th grade enrollment occurred if the number of 12th graders was either missing, larger than the total enrollment, or less than half of the total completers (the sum of four fields: regular diplomas, plus other diplomas, plus other high school completers, plus high school equivalencies).

If the grade span was reasonable, the value of the total enrollment divided by the number of grades was used for 12th grade enrollment. Otherwise, if there was a 12th grade and the number of completers was greater than zero, the grade 12 enrollment was set equal to the completers. If 12th grade was not offered or the number of completers was zero, count of 12th graders was imputed to be zero.

A small number of erroneous values for high grade in 1986-87 were set to 12. These were cases in which there were 12th graders enrolled and completers the next year but for which high grade was less than 12. Counts of completers were transformed to ratios to preceding years' 12th graders.

PROC IMPUTE was run after the file was prepared. Variables included were average ethnic percentages and percent in poverty, as well as the average over years of each of the four categories of completers. The latter averages, which normally would be no greater than 1, unless

there was substantial in-migration, were not allowed to exceed 2. Values of percentage of 12th graders who earned regular diplomas that differed from the average (across years) by more than 50 percentage points and value-, of other completion types that differed by more than 20 percentage points from the average were set to missing. Hot deck distributions were selected separately for three sizes of 12th grade cohorts: <20, 20 to 99, and 100 or more. The results were transformed back to counts, and three districts new in 1991-92 were separately imputed to have no completers.

Step 14. All imputed counts on the file were rounded to integers.

Step 15. Impute Per-Pupil Revenues and Expenditures. In addition to variables on the CCD Nonfiscal survey file, two variables on the F-33 Census of Governments survey, total revenues and expenditures per pupil, were imputed for the four school years from 1989-90 through 1992-3. For nearly every regular school district, data were present for at least one of the four years. Districts with data in none of these, years (n = 90) were imputed as the average value of per-pupil *revenues* and expenditures for districts reporting data in the specified year, by category. The categories for which separate mean values were computed in each of the four year,-, were large and small districts in rural and nonrural settings in each of the four standard geographic regions. (The division of the south into two subregions used elsewhere in this report was not applied to this imputation.)

For all imputations, the first step was to compute mean values of per-pupil revenues and expenditures for the 1 1,729 regular districts with F-33 data in all four of the school years (1 989-90 through 1992-93). The mean values for per-pupil revenues and expenditures were obtained for each of four regions, separately for small and large rural and nonrural districts in each year (a total of 129 numbers). Means were weighted by the F-33 estimate of enrollment in the year.

Next, for each pair of adjacent years, a linear regression function was estimated, using a single predictor (the same measure in the adjacent year), to predict. the deviation of a district's per-pupil revenues or expenditures from the mean for that district's region and size and locale category. A total of 12 regressions were estimated (3 pairs of adjacent years, in each order, for revenues and expenditures). The regressions were weighted by the F-33 estimate of enrollment in the year being predicted. Then, for cases missing in a year, the value was imputed as the sum of (a) the mean value for the region by size by locale category for that year and (b) the estimated deviation from the mean based on the regression.

The percentages of data that were imputed for this report range from 0.0 percent to 47.7 percent, as shown in table B2. Except for race and special education counts in the *earlier years, none* of these percentages were as great as 20 percent. Although these percentages primarily represent missing data, some imputed values are the result of setting unreasonable reported values to missing. As a general rule, most imputed values were based on reported values for the same district in different years, using the rules summarized above. It should be noted that these percentages pertain only to regular school districts, as used in this report. Between 1,000 and 2,000 other entities are included in the Common Core of Data public school district release file.

APPENDIX C

SAMPLE COGNOS SCREEN SHOTS

File Edit View Explore Format Tools Window Help

Year | Region | Typecode Description | Location | Metro Status Description | Enrollment Category | Grade Span Category | Agency Count

	West	South	Northeast	Midwest	Region
1986	2995	3401	2984	5981	15361
1987	2989	3384	2983	5930	15286
1988	2979	3379	2973	5878	15209
1989	2958	3365	2968	5830	15121
1990	2946	3348	2971	5780	15045
1991	2940	3324	2969	5737	14970
1992	2918	3305	2969	5645	14837
1993	2879	3268	2961	5556	14684
Year	23604	26774	23778	46337	120493

For Help, press F1

BEST COPY AVAILABLE

Year | Northeast | Typecode Description | Location | Metro Status Description | Enrollment Category | Grade Span Category | Total Schools

Rural

Layer 1 of 8

	CT	MA	ME	NH	NJ	NY	PA	RI	Northeast
1986	46	132	266	134	135	469	546	9	1904
1987	46	131	276	134	134	470	536	8	1902
1988	46	131	269	138	136	468	532	12	1900
1989	46	130	263	141	135	469	525	11	1891
1990	46	130	262	139	137	473	516	13	1927
1991	45	127	255	147	138	467	508	13	1968
1992	46	127	260	150	138	473	500	13	1875
1993	46	129	256	153	135	473	502	13	1874
Year	367	1037	2107	1136	1088	3762	4165	92	15141

Northeast

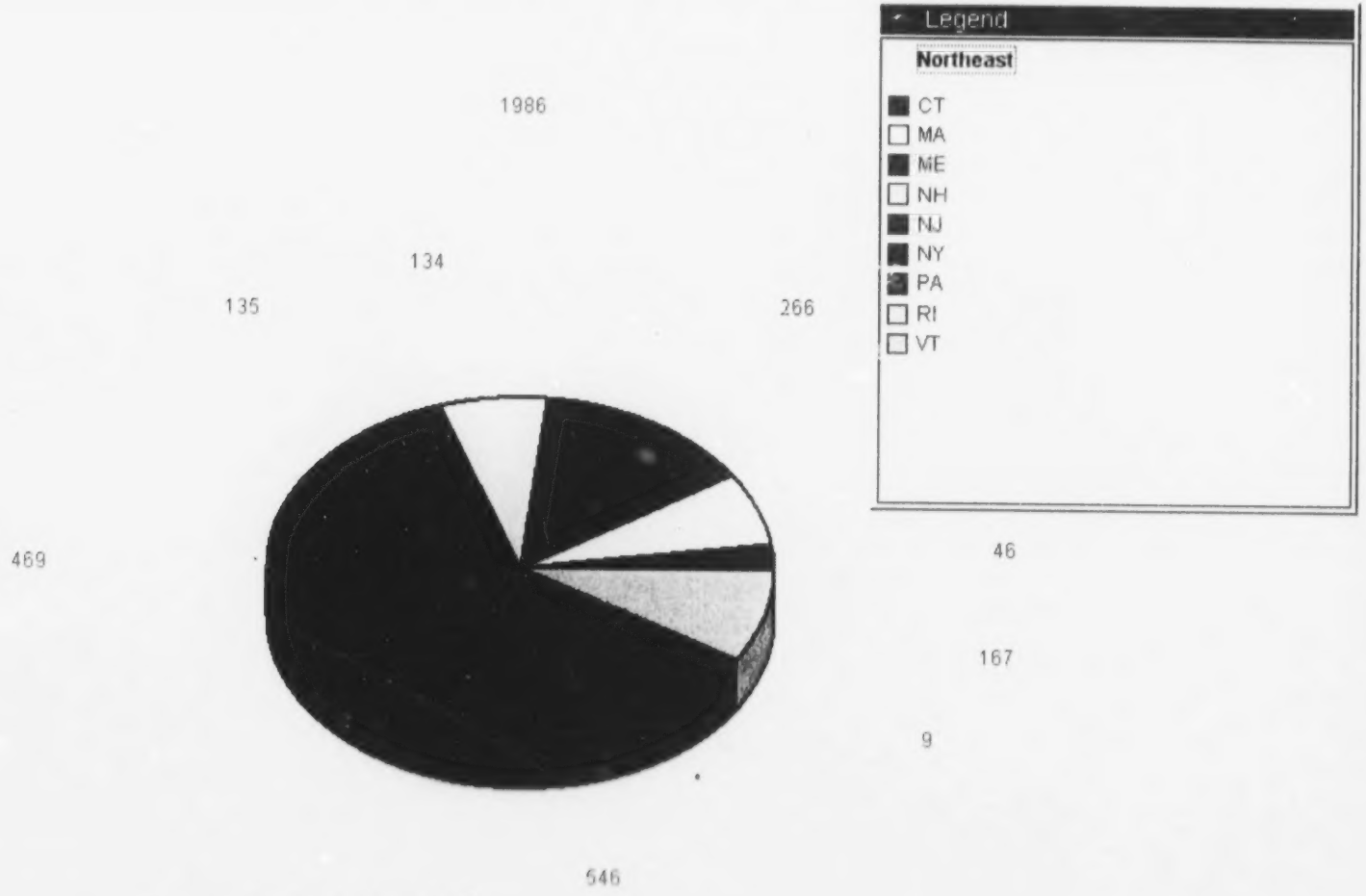
BEST COPY AVAILABLE

File Edit View Explore Format Tools Window Help

Year | Northeast | Typecode Description | Location | Metro Status Description | Enrollment Category | Grade Span Category | Total Schools

Rural

Layer 1 of 8



BEST COPY AVAILABLE

1904 (1986 Northeast)

1986

File Edit View Explore Calculate Format Tools Window Help

Year | Region | Typecode Description | Location | Metro Status Description | Enrollment Category | Grade Span Category | White

	Percent Asian	Percent Native American	Percent Hispanic	Percent Black	Percent White	Total Students
1986	0.0%	0.0%	0.0%	0.0%	0.0%	39,590,333
1987	3.0%	0.9%	10.1%	16.4%	69.5%	39,746,179
1988	3.1%	1.0%	10.6%	16.3%	69.1%	39,940,617
1989	3.2%	1.0%	11.0%	16.2%	68.6%	40,309,035
1990	3.3%	1.0%	11.5%	16.2%	68.0%	40,968,299
1991	3.4%	1.0%	11.9%	16.2%	67.5%	41,812,679
1992	3.4%	1.0%	11.9%	16.1%	67.5%	42,576,009
1993	3.7%	1.1%	13.0%	17.0%	67.7%	42,194,664
Year	2.9%	0.9%	10.1%	14.4%	60.0%	327,137,815

BEST COPY AVAILABLE

Total Students

File Edit View Explore Calculate Format Tools Window Help

Year | Region | Typecode Description | Location | Metro Status Description | Enrollment Category | Grade Span Category | White

Northeast

Layer 1 of 5

	Percent Asian	Percent Native American	Percent Hispanic	Percent Black	Percent White	Total Students
1986	0.0%	0.0%	0.0%	0.0%	0.0%	7,124,697
1987	2.8%	0.2%	9.0%	14.8%	73.2%	7,139,009
1988	2.9%	0.2%	9.1%	14.8%	73.1%	7,095,409
1989	3.1%	0.2%	9.3%	14.8%	72.6%	7,089,433
1990	3.2%	0.2%	9.7%	14.8%	72.1%	7,171,485
1991	3.3%	0.2%	10.0%	14.9%	71.6%	7,298,578
1992	3.3%	0.2%	10.0%	14.9%	71.5%	7,428,993
1993	4.1%	0.3%	12.2%	17.5%	81.3%	6,540,933
Year	2.8%	0.2%	8.6%	13.3%	64.3%	56,888,537

Total Students

BEST COPY AVAILABLE

Listing of NCES Working Papers to Date

Please contact Ruth R. Harris at (202) 219-1831 (ruth_harris@ed.gov)
if you are interested in any of the following papers

Number	Title	Contact
94-01 (July)	Schools and Staffing Survey (SASS) Papers Presented at Meetings of the American Statistical Association	Dan Kasprzyk
94-02 (July)	Generalized Variance Estimate for Schools and Staffing Survey (SASS)	Dan Kasprzyk
94-03 (July)	1991 Schools and Staffing Survey (SASS) Reinterview Response Variance Report	Dan Kasprzyk
94-04 (July)	The Accuracy of Teachers' Self-reports on their Postsecondary Education: Teacher Transcript Study, Schools and Staffing Survey	Dan Kasprzyk
94-05 (July)	Cost-of-Education Differentials Across the States	William Fowler
94-06 (July)	Six Papers on Teachers from the 1990-91 Schools and Staffing Survey and Other Related Surveys	Dan Kasprzyk
94-07 (Nov.)	Data Comparability and Public Policy: New Interest in Public Library Data Papers Presented at Meetings of the American Statistical Association	Carrol Kindel
95-01 (Jan.)	Schools and Staffing Survey: 1994 Papers Presented at the 1994 Meeting of the American Statistical Association	Dan Kasprzyk
95-02 (Jan.)	QED Estimates of the 1990-91 Schools and Staffing Survey: Deriving and Comparing QED School Estimates with CCD Estimates	Dan Kasprzyk
95-03 (Jan.)	Schools and Staffing Survey: 1990-91 SASS Cross-Questionnaire Analysis	Dan Kasprzyk
95-04 (Jan.)	National Education Longitudinal Study of 1988: Second Follow-up Questionnaire Content Areas and Research Issues	Jeffrey Owings
95-05 (Jan.)	National Education Longitudinal Study of 1988: Conducting Trend Analyses of NLS-72, HS&B, and NELS:88 Seniors	Jeffrey Owings

Listing of NCES Working Papers to Date--Continued

Number	Title	Contact
95-06 (Jan.)	National Education Longitudinal Study of 1988: Conducting Cross-Cohort Comparisons Using HS&B, NAEP, and NELS:88 Academic Transcript Data	Jeffrey Owings
95-07 (Jan.)	National Education Longitudinal Study of 1988: Conducting Trend Analyses HS&B and NELS:88 Sophomore Cohort Dropouts	Jeffrey Owings
95-08 (Feb.)	CCD Adjustment to the 1990-91 SASS: A Comparison of Estimates	Dan Kasprzyk
95-09 (Feb.)	The Results of the 1993 Teacher List Validation Study (TLVS)	Dan Kasprzyk
95-10 (Feb.)	The Results of the 1991-92 Teacher Follow-up Survey (TFS) Reinterview and Extensive Reconciliation	Dan Kasprzyk
95-11 (Mar.)	Measuring Instruction, Curriculum Content, and Instructional Resources: The Status of Recent Work	Sharon Bobbitt & John Ralph
95-12 (Mar.)	Rural Education Data User's Guide	Samuel Peng
95-13 (Mar.)	Assessing Students with Disabilities and Limited English Proficiency	James Houser
95-14 (Mar.)	Empirical Evaluation of Social, Psychological, & Educational Construct Variables Used in NCES Surveys	Samuel Peng
95-15 (Apr.)	Classroom Instructional Processes: A Review of Existing Measurement Approaches and Their Applicability for the Teacher Follow-up Survey	Sharon Bobbitt
95-16 (Apr.)	Intersurvey Consistency in NCES Private School Surveys	Steven Kaufman
95-17 (May)	Estimates of Expenditures for Private K-12 Schools	Stephen Broughman
95-18 (Nov.)	An Agenda for Research on Teachers and Schools: Revisiting NCES' Schools and Staffing Survey	Dan Kasprzyk
96-01 (Jan.)	Methodological Issues in the Study of Teachers' Careers: Critical Features of a Truly Longitudinal Study	Dan Kasprzyk

Listing of NCES Working Papers to Date--Continued

Number	Title	Contact
96-02 (Feb.)	Schools and Staffing Survey (SASS): 1995 Selected papers presented at the 1995 Meeting of the American Statistical Association	Dan Kasprzyk
96-03 (Feb.)	National Education Longitudinal Study of 1988 (NELS:88) Research Framework and Issues	Jeffrey Owings
96-04 (Feb.)	Census Mapping Project/School District Data Book	Tai Phan
96-05 (Feb.)	Cognitive Research on the Teacher Listing Form for the Schools and Staffing Survey	Dan Kasprzyk
96-06 (Mar.)	The Schools and Staffing Survey (SASS) for 1998-99: Design Recommendations to Inform Broad Education Policy	Dan Kasprzyk
96-07 (Mar.)	Should SASS Measure Instructional Processes and Teacher Effectiveness?	Dan Kasprzyk
96-08 (Apr.)	How Accurate are Teacher Judgments of Students' Academic Performance?	Jerry West
96-09 (Apr.)	Making Data Relevant for Policy Discussions: Redesigning the School Administrator Questionnaire for the 1998-99 SASS	Dan Kasprzyk
96-10 (Apr.)	1998-99 Schools and Staffing Survey: Issues Related to Survey Depth	Dan Kasprzyk
96-11 (June)	Towards an Organizational Database on America's Schools: A Proposal for the Future of SASS, with comments on School Reform, Governance, and Finance	Dan Kasprzyk
96-12 (June)	Predictors of Retention, Transfer, and Attrition of Special and General Education Teachers: Data from the 1989 Teacher Followup Survey	Dan Kasprzyk
96-13 (June)	Estimation of Response Bias in the NHES:95 Adult Education Survey	Steven Kaufman
96-14 (June)	The 1995 National Household Education Survey: Reinterview Results for the Adult Education Component	Steven Kaufman

Listing of NCES Working Papers to Date--Continued

Number	Title	Contact
96-15 (June)	Nested Structures: District-Level Data in the Schools and Staffing Survey	Dan Kasprzyk
96-16 (June)	Strategies for Collecting Finance Data from Private Schools	Stephen Broughman
96-17 (July)	National Postsecondary Student Aid Study: 1996 Field Test Methodology Report	Andrew G. Malizio
96-18 (Aug.)	Assessment of Social Competence, Adaptive Behaviors, and Approaches to Learning with Young Children	Jerry West
96-19 (Oct.)	Assessment and Analysis of School-Level Expenditures	William Fowler
96-20 (Oct.)	1991 National Household Education Survey (NHES:91) Questionnaires: Screener, Early Childhood Education, and Adult Education	Kathryn Chandler
96-21 (Oct.)	1993 National Household Education Survey (NHES:93) Questionnaires: Screener, School Readiness, and School Safety and Discipline	Kathryn Chandler
96-22 (Oct.)	1995 National Household Education Survey (NHES:95) Questionnaires: Screener, Early Childhood Program Participation, and Adult Education	Kathryn Chandler
96-23 (Oct.)	Linking Student Data to SASS: Why, When, How	Dan Kasprzyk
96-24 (Oct.)	National Assessments of Teacher Quality	Dan Kasprzyk
96-25 (Oct.)	Measures of Inservice Professional Development: Suggested Items for the 1998-1999 Schools and Staffing Survey	Dan Kasprzyk
96-26 (Nov.)	Improving the Coverage of Private Elementary-Secondary Schools	Steven Kaufman
96-27 (Nov.)	Intersurvey Consistency in NCES Private School Surveys for 1993-94	Steven Kaufman

Listing of NCES Working Papers to Date--Continued

Number	Title	Contact
96-28 (Nov.)	Student Learning, Teaching Quality, and Professional Development: Theoretical Linkages, Current Measurement, and Recommendations for Future Data Collection	Mary Rollefson
96-29 (Nov.)	Undercoverage Bias in Estimates of Characteristics of Adults and 0- to 2-Year-Olds in the 1995 National Household Education Survey (NHES:95)	Kathryn Chandler
96-30 (Dec.)	Comparison of Estimates from the 1995 National Household Education Survey (NHES:95)	Kathryn Chandler
97-01 (Feb.)	Selected Papers on Education Surveys: Papers Presented at the 1996 Meeting of the American Statistical Association	Dan Kasprzyk
97-02 (Feb.)	Telephone Coverage Bias and Recorded Interviews in the 1993 National Household Education Survey (NHES:93)	Kathryn Chandler
97-03 (Feb.)	1991 and 1995 National Household Education Survey Questionnaires: NHES:91 Screener, NHES:91 Adult Education, NHES:95 Basic Screener, and NHES:95 Adult Education	Kathryn Chandler
97-04 (Feb.)	Design, Data Collection, Monitoring, Interview Administration Time, and Data Editing in the 1993 National Household Education Survey (NHES:93)	Kathryn Chandler
97-05 (Feb.)	Unit and Item Response, Weighting, and Imputation Procedures in the 1993 National Household Education Survey (NHES:93)	Kathryn Chandler
97-06 (Feb.)	Unit and Item Response, Weighting, and Imputation Procedures in the 1995 National Household Education Survey (NHES:95)	Kathryn Chandler
97-07 (Mar.)	The Determinants of Per-Pupil Expenditures in Private Elementary and Secondary Schools: An Exploratory Analysis	Stephen Broughman
97-08 (Mar.)	Design, Data Collection, Interview Timing, and Data Editing in the 1995 National Household Education Survey	Kathryn Chandler

Listing of NCES Working Papers to Date--Continued

Number	Title	Contact
97-09 (Apr.)	Status of Data on Crime and Violence in Schools: Final Report	Lee Hoffman
97-10 (Apr.)	Report of Cognitive Research on the Public and Private School Teacher Questionnaires for the Schools and Staffing Survey 1993-94 School Year	Dan Kasprzyk
97-11 (Apr.)	International Comparisons of Inservice Professional Development	Dan Kasprzyk
97-12 (Apr.)	Measuring School Reform: Recommendations for Future SASS Data Collection	Mary Rollefson
97-13 (Apr.)	Improving Data Quality in NCES: Database-to-Report Process	Susan Ahmed
97-14 (Apr.)	Optimal Choice of Periodicities for the Schools and Staffing Survey: Modeling and Analysis	Steven Kaufman
97-15 (May)	Customer Service Survey: Common Core of Data Coordinators	Lee Hoffman
97-16 (May)	International Education Expenditure Comparability Study: Final Report, Volume I	Shelley Burns
97-17 (May)	International Education Expenditure Comparability Study: Final Report, Volume II, Quantitative Analysis of Expenditure Comparability	Shelley Burns
97-18 (June)	Improving the Mail Return Rates of SASS Surveys: A Review of the Literature	Steven Kaufman
97-19 (June)	National Household Education Survey of 1995: Adult Education Course Coding Manual	Peter Stowe
97-20 (June)	National Household Education Survey of 1995: Adult Education Course Code Merge Files User's Guide	Peter Stowe
97-21 (June)	Statistics for Policymakers or Everything You Wanted to Know About Statistics But Thought You Could Never Understand	Susan Ahmed
97-22 (July)	Collection of Private School Finance Data: Development of a Questionnaire	Stephen Broughman

Listing of NCES Working Papers to Date--Continued

Number	Title	Contact
97-23 (July)	Further Cognitive Research on the Schools and Staffing Survey (SASS) Teacher Listing Form	Dan Kasprzyk
97-24 (Aug.)	Formulating a Design for the ECLS: A Review of Longitudinal Studies	Jerry West
97-25 (Aug.)	1996 National Household Education Survey (NHES:96) Questionnaires: Screener/Household and Library, Parent and Family Involvement in Education and Civic Involvement, Youth Civic Involvement, and Adult Civic Involvement	Kathryn Chandler
97-26 (Oct.)	Strategies for Improving Accuracy of Postsecondary Faculty Lists	Linda Zimbler
97-27 (Oct.)	Pilot Test of IPEDS Finance Survey	Peter Stowe
97-28 (Oct.)	Comparison of Estimates in the 1996 National Household Education Survey	Kathryn Chandler
97-29 (Oct.)	Can State Assessment Data be Used to Reduce State NAEP Sample Sizes?	Steven Gorman
97-30 (Oct.)	ACT's NAEP Redesign Project: Assessment Design is the Key to Useful and Stable Assessment Results	Steven Gorman
97-31 (Oct.)	NAEP Reconfigured: An Integrated Redesign of the National Assessment of Educational Progress	Steven Gorman
97-32 (Oct.)	Innovative Solutions to Intractable Large Scale Assessment (Problem 2: Background Questionnaires)	Steven Gorman
97-33 (Oct.)	Adult Literacy: An International Perspective	Marilyn Binkley
97-34 (Oct.)	Comparison of Estimates from the 1993 National Household Education Survey	Kathryn Chandler
97-35 (Oct.)	Design, Data Collection, Interview Administration Time, and Data Editing in the 1996 National Household Education Survey	Kathryn Chandler
97-36 (Oct.)	Measuring the Quality of Program Environments in Head Start and Other Early Childhood Programs: A Review and Recommendations for Future Research	Jerry West

Listing of NCES Working Papers to Date--Continued

Number	Title	Contact
97-37 (Nov.)	Optimal Rating Procedures and Methodology for NAEP Open-ended Items	Steven Gorman
97-38 (Nov.)	Reinterview Results for the Parent and Youth Components of the 1996 National Household Education Survey	Kathryn Chandler
97-39 (Nov.)	Undercoverage Bias in Estimates of Characteristics of Households and Adults in the 1996 National Household Education Survey	Kathryn Chandler
97-40 (Nov.)	Unit and Item Response Rates, Weighting, and Imputation Procedures in the 1996 National Household Education Survey	Kathryn Chandler
97-41 (Dec.)	Selected Papers on the Schools and Staffing Survey: Papers Presented at the 1997 Meeting of the American Statistical Association	Steve Kaufman
97-42 (Jan. 1998)	Improving the Measurement of Staffing Resources at the School Level: The Development of Recommendations for NCES for the Schools and Staffing Survey (SASS)	Mary Rollefson
97-43 (Dec.)	Measuring Inflation in Public School Costs	William J. Fowler, Jr.
97-44 (Dec.)	Development of a SASS 1993-94 School-Level Student Achievement Subfile: Using State Assessments and State NAEP, Feasibility Study	Michael Ross
98-01 (Jan.)	Collection of Public School Expenditure Data: Development of a Questionnaire	Stephen Broughman
98-02 (Jan.)	Response Variance in the 1993-94 Schools and Staffing Survey: A Reinterview Report	Steven Kaufman
98-03 (Feb.)	Adult Education in the 1990s: A Report on the 1991 National Household Education Survey	Peter Stowe
98-04 (Feb.)	Geographic Variations in Public Schools' Costs	William J. Fowler, Jr.

Listing of NCES Working Papers to Date--Continued

Number	Title	Contact
98-05 (Mar.)	SASS Documentation: 1993-94 SASS Student Sampling Problems; Solutions for Determining the Numerators for the SASS Private School (3B) Second-Stage Factors	Steven Kaufman
98-06 (May)	National Education Longitudinal Study of 1988 (NELS:88) Base Year through Second Follow-Up: Final Methodology Report	Ralph Lee
98-07 (May)	Decennial Census School District Project Planning Report	Tai Phan
98-08 (July)	The Redesign of the Schools and Staffing Survey for 1999-2000: A Position Paper	Dan Kasprzyk
98-09 (Aug.)	High School Curriculum Structure: Effects on Coursetaking and Achievement in Mathematics for High School Graduates—An Examination of Data from the National Education Longitudinal Study of 1988	Jeffrey Owings
98-10 (Aug.)	Adult Education Participation Decisions and Barriers: Review of Conceptual Frameworks and Empirical Studies	Peter Stowe
98-11 (Aug.)	Beginning Postsecondary Students Longitudinal Study First Follow-up (BPS:96-98) Field Test Report	Aurora D'Amico
98-12 (Oct.)	A Bootstrap Variance Estimator for Systematic PPS Sampling	Steven Kaufman
98-13 (Oct.)	Response Variance in the 1994-95 Teacher Follow-up Survey	Steven Kaufman
98-14 (Oct.)	Variance Estimation of Imputed Survey Data	Steven Kaufman
98-15 (Oct.)	Development of a Prototype System for Accessing Linked NCES Data	Steven Kaufman

DEPARTMENT OF EDUCATION
UNITED STATES OF AMERICA